U.S. DEPARTMENT OF COMMERCE
National Technical Information Service

PB-292 948

The Spatial Distribution of Cancer Mortality and of High and Low Risk Factors in the New Jersey-New York-Philadelphia Metropolitan Regions, 1950-1969. Part I

Livingston Coll, New Brunswick, NJ

Prepared for

New Jersey Dept of Environmental Protection, Trenton

Jan 79

50272-101

REPORT DOCUMENTATION PAGE

1. REPORT NO. NJDEP-79/01

2.

PB292948

4. Title and Subtitle
Cancer Mortality in the New Jersey Region, 1950-1969:
High and Low Risk Factors.
Part I.

5. Report Date 1/79

6.

7. Author(s) Michael R. Greenberg

8. Performing Organization Rept. No. NJDEP-79/01

9. Performing Organization Name and Address
Dept of Community Development, Urban Planning and Geography
Livingston College
Rutgers University, New Brunswick, N.J.

10. Project/Task/Work Unit No.

11. Contract(C) or Grant(G) No.
(C)
(G)

12. Sponsoring Organization Name and Address
New Jersey Dept. of Environ. Prot.
Prog. on Env. Cancer & Toxic Sub.
P.O. Box 1390
Trenton, N.J. 08625

13. Type of Report & Period Covered

14.

15. Supplementary Notes

16. Abstract (Limit: 200 words)

The report clearly demonstrates that there are identifiable geographic patterns of cancer mortality in the 49 counties comprising the New York-New Jersey-Philadelphia metropolitan region. This indicates that specific factors are involved, which are potentially identifiable through further research. Many such factors were investigated, and a number of then, related to both high and low risk, were evaluated and described.

There are some factors which appear more often than others: degraded air quality, ethnic background, the presence of workers in certain industries, socioeconomic level, the source of drinking water, motor vehicle traffic and diet, in particular. Because of the fact that the data used were available only at the county level, and because some of the date were more complete and sound than others, the strength of the statistical correlations derived from the data varied.

In brief, the study does not point to any one factor or group of factors as being "responsible" for the high cancer mortality rates in this region. Rather, it presents a number of high and low risk factors, and points to those which need to be evaluated in much greater detail.

17. Document Analysis a. Descriptors

b. Identifiers/Open-Ended Terms

New Jersey	Occupation	Socioeconomic Status
Cancer	Air Quality	Factor Analysis
Mortality Rate	Water Supply	Regression
Epidemiology	Ethnicity	

c. COSATI Field/Group

18. Availability Statement
Release unlimited

19. Security Class (This Report) unclassified

20. Security Class (This Page) unclassifed

21. No. of Pages 235

22. Price A11- A01

(See ANSI-Z39.18) *See Instructions on Reverse* OPTIONAL FORM 272 (4-77)
(Formerly NTIS-35)
Department of Commerce

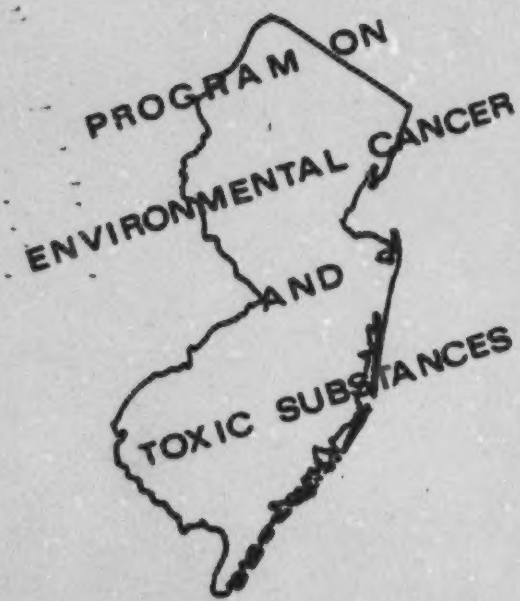
Let's protect our earth

JANUARY, 1979

PB 292948

CANCER MORTALITY IN THE NEW JERSEY REGION, 1950-1969:

HIGH AND LOW RISK FACTORS


PROGRAM ON
ENVIRONMENTAL CANCER
AND
TOXIC SUBSTANCES

STATE OF NEW JERSEY DEPARTMENT OF ENVIRONMENTAL PROTECTION

BRENDAN BYRNE
GOVERNOR

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

DANIEL J. O'HERN
COMMISSIONER

NEW JERSEY DEPARTMENT OF ENVIRONMENTAL PROTECTION

PROGRAM ON ENVIRONMENTAL CANCER AND TOXIC SUBSTANCES

THE SPACIAL DISTRIBUTION OF CANCER MORTALITY AND OF HIGH AND LOW RISK FACTORS IN THE NEW JERSEY-NEW YORK-PHILADELPHIA METROPOLITAN REGIONS, 1950-1969. PART I.

BY MICHAEL R. GREENBERG

JANUARY 1979

GLENN PAULSON, PH.D.
ASSISTANT COMMISSIONER
FOR SCIENCE

PETER W. PREUSS, PH.D.
SPECIAL ASSISTANT TO
THE COMMISSIONER

i.a

THE SPATIAL DISTRIBUTION OF CANCER MORTALITY AND OF HIGH AND LOW RISK FACTORS IN THE NEW JERSEY-NEW YORK-PHILADELPHIA METROPOLITAN REGIONS, 1950-1969

PART I

by: Michael R. Greenberg
Departments of Community Development,
Urban Planning, and Geography
Rutgers University
January 1979

i.b

DISCLAIMER

This report has been reviewed by the Program on Environmental Cancer and Toxic Substances of the New Jersey Department of Environmental Protection and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the New Jersey Department of Environmental Protection.

This study was partially funded by the United States Environmental Protection Agency (EPA grant number R 805526010).

TABLE OF CONTENTS

PAGE

Summary.. i

Chapter 1: Defining the Spatial Distribution of Cancer and of High and Low Risk Factors.....................................1

Choice of Study Area and a Cancer Data Set.....................1

An Overview of the Mathematical Techniques.... 1

A Description of the Data..................................19

Summary...64

Chapter 2: Spatial Association of Cancer Mortalities in the New Jersey-New York-Philadelphia Metropolitan Areas..................65

Selection of Data..................................65

Cancer Mortality: Comparison of the Region and the Nation.....68

Description of Intraregional Variation........................70

Summary...107

Chapter 3: Screening of High and Low Risk Indicators..............108

Personal Environment Risk Factors...........................125

Occupational Environment Risk Factors.........................128

Local Ambient Environment Risk Factors........................130

Factor Analysis Indicators..................................133

Expected and Unexpected Correlations: An Overview............137

Summary...140

Chapter 4: Results by Disease

White Male Prostate Cancer Mortality..........................142

White Female and Male Bladder Cancer Mortality................142

White Female and Male Skin Cancer Mortality...................148

PAGE

White Female and Male Leukemia Cancer Mortality..............151

White Female and Male Kidney Cancer Mortality................154

White Female and Male Pancreas Cancer Mortality..............157

White Female Breast Cancer Mortality.........................160

White Female and Male Lung Cancer Mortality..................164

White Female and Male Large Intestine Cancer Mortality........171

White Female and Male Liver Cancer Mortality.................174

White Female and Male Rectum Cancer Mortality................179

White Female and Male Stomach Cancer Mortality...............181

White Female and Male All Cancer Mortality...................184

Urban/Industrial Cancer Corridor Factor......................188

The Southern New Jersey Factor...............................189

Northeastern Pennsylvania Female Cancer Mortality Factor......190

Summary..191

Citations..197

Appendix...200

Figures

Number		Page
S-1	Forty-Nine County Study Area	ix
1-1	Association of Female White Stomach Cancer Rates and Polish Foreign Born Population in New Jersey, 1950-69	3
2-1	Forty-Nine County Study Area	66
2-2	Male White Cancer Mortality, 1950-69	71
2-3	Female White Cancer Mortality, 1950-69	72
2-4	Male Nonwhite Cancer Mortality, 1950-69	73
2-5	Female Nonwhite Cancer Mortality, 1950-69	74
2-6	Male White Stomach Cancer Mortality, 1950-69	76
2-7	Female White Stomach Cancer Mortality, 1950-60	77
2-8	Male White Large Intestine Mortality, 1950-69	78
2-9	Female White Large Intestine Mortality, 1950-69	79
2-10	Male White Lung Cancer Mortality, 1950-69	81
2-11	Female White Lung Cancer Mortality, 1950-69	82
2-12	Female White Breast Cancer Mortality, 1950-69	83
2-13	Male White Bladder Cancer Mortality, 1950-69	84
2-14	Female White Bladder Cancer Mortality, 1950-69	85
2-15	Male White Skin Cancer Mortality, 1950-69	86
2-16	Male White Leukemia Mortality, 1950-69	87
2-17	White Population Composite Cancer Mortality Groups, 1950-69	95
2-18	Nonwhite Population Composite Mortality Groups, 1950-69	103
3-1	Composite Risk Groups	138

Tables

Number Page

1- Cancer Sites Mapped and Tabulated at the County Scale by the National Cancer Institite 1950-1969....21

2- Summary of Strongest High Risk Cancer Factors....23

3- Expected Association between Occupation Variables and Cancer....37

4- Location of Air Monitoring Stations for Which We Collected Data 1953-1957, 1972-1975....43

5- TSP Air Quality Indicator Average Concentration ug/m^3....48

6- Particulate Emissions 1970....50

7- Counties by Wind Station....52

8- Air Station Data....52

9- Forty-Nine County Study Area....67

10- Cancer Mortalities by Body Site per 100,000 National & Regional Averages, 1950-1969....68

11- Cancer Mortality by Body Site: Ratio of Regional To National Average, 1950-1969....69

12- Interrelationship of Male-Female and White-Nonwhite Cancer Mortality Rates....88

13- Bivariate Correlations ≥ 0.5 among Thirteen White Population Cancer Mortalities over 49 County Study Area....89

14- Bivariate Correlations ≥ 0.5 among Thirteen White Population Cancer Mortalities Summarized by Body Site....90

15- Nonwhite Cancer Mortalities in Ten Counties Eliminated from the Statistical Analysis....91

16- Bivariate Correlations ≥ 0.5 among Thirteen Nonwhite Population Cancer Mortalities over 35 Political Unit Study Area....92

17- Bivariate Correlations ≥ 0.5 among Thirteen Nonwhite Population Cancer Mortalities Summarized by Body Site....93

18- White Population Urban-Industrial Cancer Corridor Factor....96

19- Counties Belonging to the Urban-Industrial Cancer Corridor Group....97

20- Counties Belonging to the Low Cancer Mortality Group....98

Page

21- Comparison of Cancer Mortality Rates in Counties with Extremely Different Rates....99

22- White Population Southern New Jersey Bladder, Female Large Intestine, and Male Leukemia Factor....100

23- Counties Belonging to the Southern New Jersey Bladder, Female Large Intestine, and Male Leukemia Factor....100

24- Northeastern Pennsylvania White Female Population Cancer Factor.101

25- Counties Belonging to Northern Pennsylvania Female Cancer Factor....102

26- New Jersey Corridor High Nonwhite Cancer Mortality Group Factor.104

27- Counties Belonging to New Jersey Corridor High Nonwhite Cancer Mortality Group....104

28- Counties Belonging to Nonwhite Low Mortality Group....106

29- Central Pennsylvania & Northeastern New Jersey Nonwhite Male Liver, Rectum, and Kidney Cancer Factor....106

30- Counties Belonging to the Central Pennsylvania and Northeastern New Jersey Nonwhite Male Liver, Rectum and Kidney Cancer Factor.107

31- Bivariate Correlations: 24 Body Sites and 62 High Risk Indicators....109

32- Screening of Personal Environment Factors....125

33- Comparison of Occupational Mix and Local Ambient Environments of 4 Ethnic Groups....127

34- Occupational Variables Retained for Final Correlations....129

35- Test of Air Emissions/Wind Direction/Distance Indicators Correlations with Cancer Mortality....132

36- Urban/Industrial Core Factor....134

37- Counties Most Strongly Identifying with the Urban/Industrial Core Factor....135

38- Italian/Polish Corridor Factor....135

39- Counties Most Strongly Identifying with the Polish-Italian Corridor County Factor....136

40- Socioeconomic Status Factor....137

41- Counties Strongly Identifying with the Socioeconomic Status Factor....137

Page

42- Pattern of Correlations Classified as Unexpected 139

43- Significant Bivariate Correlations: White Male Prostate Cancer Mortality, 1950-1969.....142

44- Correlation Analysis, White Female Bladder Cancer Mortality 1950-1969.....143

45- Significant Bivariate Correlations: White Male Bladder Cancer Mortality, 1950-1969.....145

46- Multivariate Analysis of White Male Bladder Cancer Mortality, 1950-1969, 45, 44, and 40 Area Regions.....146

47- Multivariate Analysis of White Male Bladder Cancer Mortality, 1950-1969, 21 and 20 Area Regions.....146

48- Comparison of Proportion of Chemical Workers and Male White Bladder Cancer.....147

49- Significant Bivariate Correlations: Female White Skin Cancer Mortality, 1950-1969.....149

50- Multivariate Analysis of White Female Skin Cancer Mortality, 1950-1969.....149

51- Significant Bivariate Correlations: White Male Skin Cancer Mortality, 1950-1969.....150

52- Multivariate Analysis of White Male Skin Cancer Mortality, 1950-1969.....151

53- Significant Bivariate and Multivariate Correlations: White Female Leukemia Cancer Mortality, 1950-1969.....152

54- Important Bivariate Correlations: White Male Leukemia Cancer Mortality, 1950-1969.....153

55- Multivariate Analysis of White Male Leukemia Cancer Mortality, 1950-1969.....154

56- Bivariate and Multivariate Analysis of White Female Kidney Cancer Mortality, 1950-1969.....155

57- Important Bivariate Correlations: White Male Kidney Cancer Mortality: 1950-1969.....156

58- Multivariate Analysis of White Female Kidney Cancer Mortality, 1950-1969.....157

Page

59- Important Bivariate Correlations of White Female Pancreas Cancer Mortality, 1950-1969 158

60- Multivariate Analysis of White Female Pancreas Cancer Mortality, 1950-1969 158

61- Important Bivariate Correlations: White Male Pancreas Cancer Mortality: 1950-1969 159

62- Multivariate Analysis of White Male Pancreas Cancer Mortality, 1950-1969 160

63- Important Bivariate Correlations: White Female Breast Cancer Mortality, 1950-1969 161

64- Multivariate Analysis of White Female Breast Cancer Mortality, 1950-1969, Socioeconomic Status and German Ethnic 162

65- Multivariate Analysis of White Female Breast Cancer Mortality, 1950-1969, Socioeconomic Status and Ethnic Factors 162

66- Multivariate Analysis of White Female Breast Cancer Mortality 1950-69, All Significant Indicators 163

67- Multivariate Analysis of White Female Breast Cancer Mortality, 1950-1969, Factor Analysis Variables 164

68- Important Bivariate Correlations: White Female Lung Cancer Mortality, 1950-1969 165

69- Multivariate Analysis of White Female Lung Cancer Mortality, 1950-1969 166

70- Important Bivariate Correlations: White Male Lung Cancer, 1950-1969 167

71- Correlations of Selected High Risk Indicators, 45 Area Region ... 169

72- Multivariate Analysis of White Male Lung Cancer Mortality, 1950-1969 170

73- Important Bivariate Correlations: White Female Large Intestine Mortality, 1950-1969 172

74- Important Bivariate Correlations, White Male Large Intestine Cancer Mortality, 1950-1969 173

75- Multivariate Analysis of White Male Large Intestine Cancer Mortality, 1950-1969 173

Page

76- Important Bivariate Correlations: White Female Liver Cancer Mortality, 1950-1969....175

77- Multivariate Analysis of White Female Liver Cancer Mortality, 1950-1969....175

78- Important Bivariate Correlations: White Male Liver Cancer Mortality, 1950-1969....177

79- Multivariate Analysis of White Male Liver Cancer Mortality, 1950-1969....178

80- Important Bivariate Correlations: White Female & Male Rectum Cancer Mortality....180

81- Multivariate Analysis of White Female & Male Rectum Cancer Mortality, 1950-1969....180

82- Important Bivariate Correlations: White Female & Male Stomach Cancer Mortality, 1950-1969....182

83- Multivariate Analysis of White Female & Stomach Cancer Mortality, 1950-1969....183

84- Important Bivariate Correlations, White Female and Male All Cancer Mortality, 1950-1969....185

85- Multivariate Analysis of White Female & Male All Cancer Mortality, 1950-1969....187

86- Urban-industrial Cancer Corridor Factor....188

87- Counties Belonging to Urban/Industrial Cancer Corridor Group....188

88- Multivariate Analysis of Urban/Industrial Cancer Corridor Factor....189

89- White Population Southern New Jersey Cancer Factor....189

90- Factor Scores for Counties Strongly Identifying with Southern New Jersey Factor....189

91- Important Bivariate Correlations: Southern New Jersey Factor...190

92- Northern Pennsylvania Cancer Factor....190

93- Factor Sactor Scores for Counties Strongly Identifying with the Northeastern Pennsylvania Cancer Factor....191

94- Important Bivariate Correlations: Northeastern Pennsylvania Cancer Factor....191

Page

95- Summary of Statistical Results by Body Site....................192

Appendix A - Statistical Information about the White Population, 45 Unit Study Area...200

B - Statistical Information about the Nonwhite Population, 35 Unit Study Area...205

Acknowledgements

While I am responsible for the content of this report, it could not have been completed without the contributions of many people. Peter Preuss, Glenn Paulson, Thomas Burke, Sidney Gray, Judy Louis and Paul White of the New Jersey Department of Environmental Protection saw the need for and actively encouraged us to pursue the work by supplying funds, data, and suggestions. My graduate students patiently gathered data and began the analysis of the results. The following major contributions were made. Warren Feld, Beth Krugman and William Spencer studied and gathered cancer mortality and incidence data. Ed Frankel and Barie Kline secured personal environmental data. Richard Anderson, Ken Hess, Debbie Malek, Jay Marlles and Eric Synder obtained the huge occupational data set. Ken Gallagher and Bill Page gathered the air pollution information. Chris Demus and Andy Blalock obtained the transportation data. Katherine McVicker gathered the fossil fuel generating station data. Debbie Malek, Francis Minchella, and Mark Usinger focused on water supply information. Rebecca King provided a legal perspective. John Grimaldi made computer maps.

Special acknowledgement must go to Richard Anderson and John Caruana who organized the data and performed numerous statistical analyses. Finally, I would like to thank Gwen Suggs for drawing the maps and Vera Lee who once again had to read and type from over a thousand pages of my handwriting.

Michael Greenberg
January 1979

SUMMARY

In 1976, citizens of New Jersey were dismayed to learn that during the 1950's and 1960's the state had the highest white male cancer mortality rate and among the highest white female and nonwhite cancer mortality rates in the United States. The state government responded by allocating personnel and financial resources to try to determine the reasons for its unenviable position. Long before any formal research could be initiated, many people were convinced that they knew why New Jersey's rates were so high. Some have argued that polluted water and air are the causes; others contend that occupational exposures are the chief cause; still others have blamed eating, drinking, and smoking habits; a few pointed to the presence of people born in other countries; some argued that New Jersey's high cancer rates were due to the migration of former New York City and Philadelphia residents to New Jersey; and others put forth additional factors.

This plethora of reasons was joined by a plethora of approaches for testing these hypothesized explanations. Many people have argued that the state should focus on those diseases that manifest much higher rates in New Jersey than the remainder of the nation. Others disagreed, suggesting that the cancer diseases that cause the largest proportion of mortalities should be studied. Other ideas include, but are not limited to, study the factories, the water, the air, and the food. While many of these approaches are appropriate and are being initiated, each approach in and by itself could command the major proportion of the state's resources.

Before committing sizeable amounts of resources to any specific project, it seemed prudent to learn as much as possible about cancer mortality in New Jersey by using already available data. The author was charged with studying the geography of cancer and of factors thought to be associated with elevated and reduced cancer risk in New Jersey and nearby counties. The most important constraint on the study was that only readily available data were to be used. We had neither the time nor the funds to go into the field and collect new data.

As you read the report, you are urged to consider and use the results in the context of the original charge and the accompanying constraints. Because of the geographical scale of the study and the poor quality of some of the data we had to use, this study must not be interpreted as a definitive statement on what causes cancer in New Jersey and environs. If the study was not intended to offer a definitive statement on cancer in New Jersey, what does it offer? The author had four major goals:

1. To compare cancer mortality rates in New Jersey and environs with the United States in order to identify diseases which are significantly in excess in the region.
2. To describe the spatial distribution of cancer mortality and high and low risk factors in the region.
3. To find disease/risk factor/place combinations which seem prime for more detailed epidemiological study. Some of these combinations might be expected from the literature, others might not.
4. To study the change in cancer mortality rates over time in order to determine if the region's patterns are significantly different from the nation's.

The first report focuses on numbers 2 and 3. Specifically, it describes the geography of cancer mortality from 1950 throughout 1969 and high and low risk factors from the late 1930's to the early 1970's. It also takes a first cut at identifying disease/risk factor/place combinations for further study.

The second part, which we hope to submit to the state in the fall of 1978, describes the geography of cancer mortality and of high and low factors from 1968 through 1972, compares the 1968-1972 pattern to the 1950-1969 distribution, and compares changes in, the region's and nation's cancer mortality rates. In addition, population imigration is studied as a confounding factor.

The third part of the study is being designed. Tentatively, we are going to compare the region's and the nation's cancer mortality rates for specific age groups during 1950-1954, 1955-1959, 1960-1964, and 1965-1969. In addition, the geography of these diseases will be analyzed.

Overall, upon completion of the three parts of the study, we hope to be able to say how the study region differs from the nation, where within the region the differences are greatest, which age groups are the most different and provide some hypotheses on why these differences have occurred.

Turning to this first report, the study area contains all of New Jersey and nearby portions of New York, Pennsylvania, Connecticut, and Delaware (see Figure 1 at the end of the Summary). The cancer mortality data are age, sex and race standardized mortality rates developed at the county scale by the National Cancer Institute for the period 1950-1969. The high risk factor data which range from the late 1930's to the early 1970's were gathered by us based on hypotheses formulated from the cancer risk literature. The report is divided into four chapters.

1. The first chapter focuses on the data and methods use to study the geography of cancer and mathematically to associate cancer mortality with high and low risk factors. Hypotheses are stated tying cancer sites to high and low risk factor data. The weaknesses of the data are described as are the impacts of poor data on the utility of analytical methods. For example, the high risk data were so uneven in quality that we could not justify the use of path analysis, an important causal analysis method. Therefore, the methods used are simple bivariate and multivariate correlation analyses. Factor analysis was used to help with the intercorrelated variable problem.

The use of these mathematical methods may lead to false positive, false negative, and confusing results. The causes of these problems are reviewed and guidelines for controlling or at least minimizing spurious results are put forward. These include drawing hypotheses from the literature, testing the results for consistency and strength, using control diseases and control independent variables, and in general recognizing the limitations of mathematical tools.

More specifically, in order to create as few false negative and false positive associations as possible, the following guidelines were followed: (1) initially stick to hypotheses based on the literature; (2) make sure that bivariate correlations are consistently significant in multiple time periods and in different regional configurations; (3) eliminate high risk factors that produce many obviously spurious correlations, even if they produce some expected correlations; (4) test more than one indicator of the same high risk factor (e.g., air emissions, fossil fuel generating, automobile miles driven) to be sure that the associations are not spurious; (5) use general control variables (e.g., population density and manufacturing workers) to screen out weakly correlated risk factors; (6) use control diseases (e.g., prostate) that should not be intercorrelated with high risk factors; (7) whenever necessary point out the potential confusion and misinterpretation caused by intercorrelated multivariate analysis results and by mistakingly attributing relationships observed at the county scale to individuals; (8) restrict the use of regression equations to analyzing residuals because the relationships implied by the equations cannot be taken at face value; and (9) point out the absence of important data sets such as smoking habits and alchohol consumption that limit the conclusions that may be drawn from the analysis. Chapter 1 is tedious, but necessary to read to understand the limitations of the study.

2. Chapter two presents the spatial distribution of cancer mortalities in the study area. The following body sites were studied for the white and nonwhite female and male populations.

(1) Stomach
(2) Large Intestine
(3) Rectum
(4) Liver
(5) Pancreas
(6) Lung
(7) Female Breast
(8) Male Prostate
(9) Kidney
(10) Bladder
(11) Skin
(12) Leukemia
(13) All Cancers

For the period 1950-1969, the region's male and female rates are compared to the national rates. The region's stomach, large intestine, rectum, lung, female breast, bladder, and all cancer rates were considerably higher than the national average.

Using mapping and mathematical methods, five groups of counties are distinguished with respect to their white population cancer mortality patterns.

(1) Counties located in the urban-industrial core.
They exhibit high rates of all, stomach, male large intestine, rectum, lung, male liver, and female breast cancer mortality. The counties that most strongly manifest these characteristics are Hudson, Essex, Middlesex, Nassau, New York City (as a single unit), and Philadelphia.

(2) Counties located in relatively rural areas.
They demonstrate low rates of all types of cancer mortality. The five best examples are Bucks, Chester, Dutchess, Lehigh and Montgomery Counties.

(3) Counties centered in southern New Jersey, but with other outliers. They exhibit high rates of bladder, male leukemia and female large intestine cancer mortalities. The counties most strongly identifying with this group are Cape May, Gloucester, Nassau, Philadelphia, Pike, and Salem Counties.

(4) Counties centered in northeastern Pennsylvania and in adjacent areas in New Jersey and New York. They exhibit high rates of female mortalities from kidney, skin, lung and leukemia. The strongest examples are Carbon, Hunterdon, Pike, Sullivan, and Sussex Counties.

(5) Eleven of the forty-nine counties did not closely identify with any of the four groups: Bergen, Burlington, Cumberland, Delaware, Fairfield, Litchfield, New Haven, Ocean, Orange, Somerset, and Ulster Counties.

The spatial distribution of nonwhite cancer mortality was also analyzed. Ten counties were eliminated because they had too few nonwhite mortalities for reliable statistical testing. Analysis of the remaining counties distinguised four groups. The patterns are not as distinct as those uncovered among the white population and many of the counties were located in more than one group. More than forty percent of the remaining counties (14 of 35) could not be classified. Of the total of 45 counties, we were able to classify only 21. The following is an overview of the four groups:

(1) Counties located in northeastern and southwestern New Jersey. They exhibit high rates of many types of cancer mortality. Six counties strongly identify with the type: Bergen, Hudson, Middlesex, Ocean, Passaic, and Somerset Counties.

(2) Counties located in New York and Pennsylvania that demonstrate low rates of all cancer mortalities. The six strongest examples are Bucks, Cumberland, Dutchess, Rockland, Suffolk, and Ulster Counties.

(3) Counties concentrated in central Pennsylvania and in northeastern New Jersey. They exhibit high rates of male liver, rectum, and kidney cancer mortality. The six best models of the group are Bucks, Hudson, Morris, Northampton, Passaic, and Union Counties.

(4) Twenty-four counties dispersed throughout the region, that did not identify with any of the three groups.

3. Chapter 3 presents the process of reducing about 1000 hypothesized high and low risk indicators to 62. Most of the 1000 indicators were developed to ensure that the correlations were consistent. For example, each manufacturing code is represented by the following 36 and other operational variables: manufacturing establishment, workers, and production workers in 1947, 1958, 1967, 1972; manufacturing establishments, workers and production workers 1947, 1958, 1967, and 1972 divided by land area; and manufacturing establishments, workers, and production workers 1947, 1958, 1967, and 1972 weighted by a distance parameter. Once we were able to conclude that a high or low risk factor was consistent and met the other tests (screening devices) set forth in Chapter 1, we then computed an average high risk indicator. In the example case above, the average of 1947, 1958, 1967, and 1972 workers would be used to represent the four years and three categories of data.

Most of the indicators we used were not transformed or were transformed into percentages. Some however, were more complex. For example, if air pollution contributes to lung and other cancers, then an index which tries to take the movement of air pollutants into account should produce a stronger association with cancer than a simpler indicator of emissions in each county. An air emissions/wind direction/distance index was developed to test this hypothesis. The emissions of every county were divided by the distance between each county. The resulting ratio for each county was multiplied by the percentage of the time during the year that the wind blows from the county to another. As chapter 3 indicates the emissions/wind direction/distance index and most of the other derived indices produced stronger associations than the untransformed risk factors. The transformed indices are abstract approximations and may not be manifestations of reality.

Fourteen of the 23 personal environmental factors passed the screening devices. They include socioeconomic status, ethnicity, stability, and family status indicators. More than 700 occupational variables were gathered and created. After screening, thirty-three were chosen for the final analysis. Twelve air pollution and water supply variables were retained; ten are air pollution indices and two are water supply indicators.

Finally, three indicators were created by factor analysis. One defines characteristics of urban/industrial core counties. The second focuses on counties in the New York/Philadelphia corridor that have a high proportion of American citizens of Polish and Italian decent. The third contrasts counties with high socioeconomic status with counties with low socioeconomic status. Since the mathematically derived factors are not correlated, they can be used without the problem of independent variable intercorrelations confusing the results.

5. The fourth chapter reviews the mathematical association between cancer mortality and high and low risk factors among the white population. The nonwhite data did not have enough cases for analysis beyond the level presented in chapter 2.

(1) High rates of stomach cancer are found in counties containing a large proportion of people from Eastern Europe, expecially Poland: Hudson, Mercer, Middlesex, New York City, Passaic, and Union counties. The low stomach cancer mortality counties are Chester and Montgomery. They have a relatively low proportion of Eastern European Americans.

(2) White male and female large intestine mortality rates are not strongly correlated with each other. The highest white male rates are found in the region's urban core areas: Camden, Essex, Hudson, New York City and Philadelphia. The highest white female rates are found in southwestern New Jersey and adjacent Pennsylvania: Burlington, Camden, Hunterdon, Philadelphia, and Pike. Furthermore, no single or set of high and low risk factors could be uniquely associated with large intestine cancer mortality. The strongest associations are with general urban/industrial characteristics such as manufacturing, air quality, water supply, and with ethnic groups (e.g., Americans of Italian decent) that tend to concentrate in urbanized counties.

(3) White male and female rectum cancer mortality rates are moderately correlated with each other, suggesting common causation, and with many of the high risk urbanization/industralization factors. So many moderately high associations

were found that it was not feasible to separate their impacts. The high rate counties are predominantly in the urban/industrial corridor: Essex, New York City, and Passaic. However, a few counties with high rates are not heavily urbanized: Monroe and Warren. Generally, the low rate counties are rural and exurban.

(4) High white female liver cancer mortality rates were expected to be and were found to be in counties marked by low socioeconomic status (e.g., Carbon and Hudson). High white male rates were found in urban/industrial counties such as Hudson, Essex, Middlesex, and New York City. The high rate counties tended to have relatively large numbers of chemical workers, relatively large populations of USSR and Italian decent, and high air pollution levels. The high risk factors are correlated. Therefore, their effects are not easily distinguished. Counties with low white male liver cancer mortality rates include rural Dutchess, Pike and Ulster Counties.

(5) Mortality caused by cancer of the pancreas was expected to be related to ethnic factors and among the male population to the presence of the chemical industry. The male population rates were found moderately associated with the ethnic groups (especially USSR and Polish). They were highest in Cape May, Nassau, Passaic, Pike and Sullivan Counties. The female rates were not as strongly correlated with the expected ethnic factors and were weakly correlated with socioeconomic status and air quality. Whereas the male rates seem to be generally explicable, the geographical pattern of female rates defy a consistent explanation based on the data gathered for this study.

(6) Lung cancer is the leading cause of male cancer mortality and an increasing cause of female mortality. Most of the high rate counties are in the densely developed urban/industrial corridor: Camden, Hudson, Middlesex, Nassau, New York City, and Philadelphia. Several high rate female counties (Monroe and Putnam) are, however, not part of the corridor. All of the low rate counties are relatively rural.

White male lung cancer mortality was strongly associated with a set of high risk factors most strongly manifested in the urban corridor: air pollution, the chemical and transportation equipment industries, and Polish and Italian ethnic populations. Female rates were less strongly identified with the urban corridor. Nevertheless, air pollution and the presence of ethnic populations were moderately associated, and female family status (divorced, married) appeared as a significant factor.

As previously noted, this study could not control for smoking habits; this is a serious deficiency. We will try to develop smoking data in follow-up studies.

(7) White female breast cancer mortality was assumed from the literature to be positively associated with socioeconomic status. It was in this study as well. Nassau, Putnam, and Westchester Counties are relatively affluent and manifested high white female breast cancer mortality rates. However, ethnicity, local ambient environmental conditions and even occupational variables, much to our surprise, were found to strongly associate with white female breast cancer mortality. This disease requires a good deal of further attention.

(8) White male prostate cancer mortality was used as a control disease. It was not expected to and did not correlate significantly with the personal, occupational, and local ambient environmental factors.

(9) Kidney cancer mortality produced among the weakest correlation results. We had expected ethnicity and perhaps local ambient environmental factors to associate. Some of the expected associations materialized. Ethnic and water supply were the strongest for both the male and female population. They were, however, weak. Therefore, we have little confidence in them. Overall, our high risk factors do not readily explain why both urban Nassau and Union Counties and rural Sullivan, Monroe and Pike Counties have high white male rates. Nor do they suggest why Sullivan and Pike have both high white male and female rates, Rockland County both low male and female rates, and Monroe County has a high male and and a low female rate.

(10) Bladder cancer is an important disease because the region's rates are much higher than the national average. The highest female rates tend to be found in and adjacent to southern New Jersey Counties: Salem, Gloucester, Camden, Atlantic, Philadelphia, and Cape May. Pike, Warren, Hudson, and Nassau Counties also have high rates. The high male rates are dispersed throughout the region and are predominantly in relatively rural counties (Salem, Cape May, Pike, and Hunterdon). The lowest rates are also found in largely rural counties. The literature suggested that socioeconomic status and ethnicity should be associated with male and female bladder cancer, and that among the male population industry should be a factor. The results weakly supported these hypotheses. Affluence was a factor, however, a weak one as was ethnicity for the female (German) and male (Italian) populations. The relative presence of the chemical industry was correlated with bladder cancer mortality. However, the association between chemical workers and bladder cancer mortality was inconsistent. It was not correlated with some regional configurations and correlated with others.

(11) Skin cancer mortality was thought to be related to outdoor work and among the male population to the presence of the chemical industry. Female rates were found to be moderately associated with socioeconomic status and agriculture. White male rates were strongly associated with the relative presence of the chemical industry. Salem, New Castle, Middlesex, and Somerset Counties are the counties that are largely responsible for the skin cancer mortality and chemical industry correlation.

(12) Leukemia is the one systemic cancer we studied. High mortality rates were found in the metropolitan core center New York City, and adjacent Hudson, Essex, and Westchester Counties and Philadelphia. However, rural Pike, Sullivan, and Sussex Counties also exhibited high rates. We had expected to find an association with the relative presence of Jewish people and in the case of the male population with the presence of the petrochemical industry. Only the moderate associations with our surrogates for Jewish population and motor vehicle density were found.

(13) While most cancer mortalities are accounted for by the above specific cancer types, all cancer mortality was analyzed to obtain a more general perspective. Philadelphia and New York exhibited high white female and male rates, while

Dutchess, Suffolk, and Bucks Counties demonstrated low white female and male rates. The highest white male counties lie along the urban/industrial corridor: New York City, Hudson, Essex, Middlesex, and Philadelphia. The highest white female counties are also generally found in the corridor but also include Monroe and Putnam Counties. Overall, the all cancer category is strongly correlated with ethnic groups, occupations, and local ambient environmental factors common to the region's urban/industrial corridor.

Summarizing, the purpose of this first report has been to provide a perspective on the geography of cancer mortality and on high and low risk factors thought to influence cancer mortality. The results suggest that the diseases of cancer are influenced by many complex interrelated factors. The study along with Parts II and III and the state's water and air research programs hopefully will suggest to public officials and private individuals where additional efforts can be most fruitfully made.

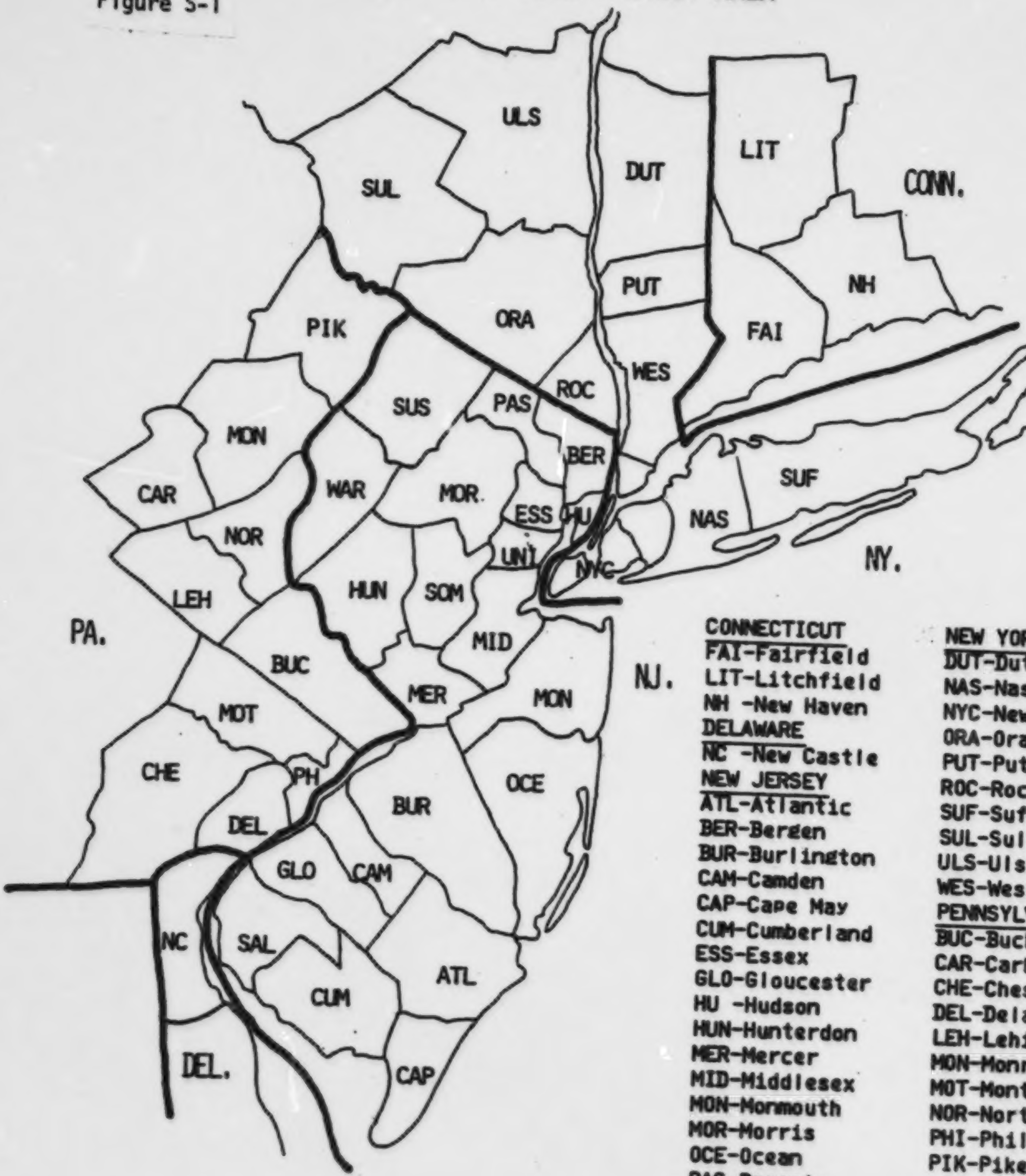
In a recent editorial in Science, Comar (June 16, 1978) using the biomedical sciences as his example, argued that

> "...all too often, published partial findings are taken uncritically at face value, misinterpreted, and misused, their qualifications are disregarded and their uncertainties forgotten. This can lead to technological fixes that do more harm than good."

This report contains descriptions of the uncertain quality of some of the data, the numerous confounding factors, methodological limitations of the research, and the fact that this is only the first of three reports. The author hopes that the reader will respect the limitations of the study and not misapply the results.

FORTY-NINE COUNTY STUDY AREA

Figure S-1


ULS
SUL
DUT
LIT
CONN.
NH
PUT
PIK
ORA
FAI
WES
ROC
SUS
PAS
MON
BER
SUF
CAR
WAR
MOR
ESS
HU
NAS
NOR
UNI
NYC
NY.
LEH
HUN
SOM
PA.
MID
BUC
MER
MON
NJ.
MOT
CHE
PH
OCE
BUR
DEL
GLO
CAM
NC
SAL
ATL
CUM
DEL.
CAP

CONNECTICUT
FAI-Fairfield
LIT-Litchfield
NH -New Haven

DELAWARE
NC -New Castle

NEW JERSEY
ATL-Atlantic
BER-Bergen
BUR-Burlington
CAM-Camden
CAP-Cape May
CUM-Cumberland
ESS-Essex
GLO-Gloucester
HU -Hudson
HUN-Hunterdon
MER-Mercer
MID-Middlesex
MON-Monmouth
MOR-Morris
OCE-Ocean
PAS-Passaic
SAL-Salem
SOM-Somerset
SUS-Sussex
UNI-Union
WAR-Warren

NEW YORK
DUT-Dutchess
NAS-Nassau
NYC-New York City
ORA-Orange
PUT-Putnam
ROC-Rockland
SUF-Suffolk
SUL-Sullivan
ULS-Ulster
WES-Westchester

PENNSYLVANIA
BUC-Bucks
CAR-Carbon
CHE-Chester
DEL-Delaware
LEH-Lehigh
MON-Monroe
MOT-Montgomery
NOR-Northampton
PHI-Philedelphia
PIK-Pike

CHAPTER 1

DEFINING THE SPATIAL DISTRIBUTION OF CANCER AND OF HIGH AND LOW RISK FACTORS

This chapter is a mathematical anaylsis of the spatial distribution of cancer and of the association between cancer and high and low risk factors. It is divided into three parts: (1) the choice of a study area and a cancer data set; (2) an overview of mathematical techniques used in the study; and (3) a description of the data.

Choice of a Study Area and a Cancer Data Set

The spatial distribution of cancer may be fruitfully studied at many geographical scales. Interesting studies range from the hamlet scale to international comparisons; and from scrutiny of a single body site to a general portrait of 30 or more body sites. Our group considered the following combinations of study areas and body sites: (1) a few body sites at the county scale for the entire Middle Atlantic region; (2) general groupings (e.g., respiratory, digestive) of body sites for 93 communities in eight New Jersey counties; and (3) the most important body sites for New Jersey and environs. Had we studied the Mid-Atlantic region, the National Cancer Institute's 1950-1969 mortality rates would have been the cancer data source. After considerable attention, the Mid-Atlantic project was rejected because preliminary investigations suggested that gathering the other necessary data would have been beyond our means.

Cancer and control disease mortality rates were collected for 93 New Jersey communities. The project has, however, been temporarily shelved because we wanted initially to study the geography of cancer over a large region.

The New Jersey and environs study area ultimately was chosen for two reasons. First, the Department of Environmental Protection of the State of New Jersey reacted enthusiastically to the proposal. Second, given the financial resources, personnel and time available, we felt that a New Jersey-centered study area would be the most feasible and useful. Once we chose the New Jersey-centered study area, the next decision was the selection of a data source for cancer information. Research and deliberations led to the conclusion that the National Cancer Institute's 1950-1969 county compilation was the only feasible set to use at this time. The NCI's data set has the major disadvantages of being a single twenty-year average and of ending in 1969. The major advantages of the NCI data are that we did not have to standardize them, the data include more than 30 body sites, and they have already been widely used and accepted as landmark information on the spatial distribution of cancer in the United States. The NCI data are described in detail later in the chapter.

An Overview of the Mathematical Techniques

Having briefly explained how a general study area and a source of cancer data were chosen, we will now review the mathematical techniques used to analyze the data.

Regression, Correlation, and Factor Analysis

It is often desirable both to observe and measure the association between two or more sets of information. Observation is done with the aid of a scatter diagram. One set of data is plotted on the vertical axis, the second is plotted on the horizontal axis. If a relationship exists, the points on the graph will follow a definite path. The following scattergram (Figure 1) shows that in New Jersey there is an association between female white stomach cancer rates and the presence of a Polish ethnic population when county scale data are used. If the association was perfect, it would form a line, rather than a path. In the sample case, all the points, except one, seem to cluster along a narrow path. The path seems to be linear. The wider the scatter of the points from a line, the less perfect the relationship between the two sets of data.

To obtain an objective assessment of the degree of association, the analyst first must choose to fit a linear or one of many non-linear lines to the data. In the sample case and in all the cases in this study, we have assumed linear associations. Linearity is a conservative assumption and probably should be assumed in the absence of information to the contrary.

Once the choice of line has been made, a mathematical process is used to find the best fitting linear line for the data. If the line is of the linear type, the equation will be of the form

$$Y_e = a + bX$$

where Y_e = the expected value of the dependent variable (the variable being predicted); a and b are constants determined by the mathematical process; and X is the value of the independent variable (the variable being used to predict the dependent variable).

In the sample case, the Y(dependent) variable is white female stomach cancer mortality, the X (independent) variable is percent Polish ethnic population at the county scale. The sample data yield the following linear equation:

$$Y_e = 6.506 + 41.630\ X$$

The equation is used to estimate an expected value of stomach cancer for each value of X. For example, in New Jersey, Salem and Hudson Counties represent the extreme white female stomach cancer rates. Salem County's actual rate is 6.6/100,000 and its Polish value is .01756 (1.756%). Substituting into the above equation, we obtain

$$Y_e\ \text{Salem} = 6.506 + 41.630\ (.01756) = 7.237$$

The actual value for Salem County is 6.6; the expected value is 7.2. For Hudson County, the actual value is 12.1/100,000, the expected value is 10.8.

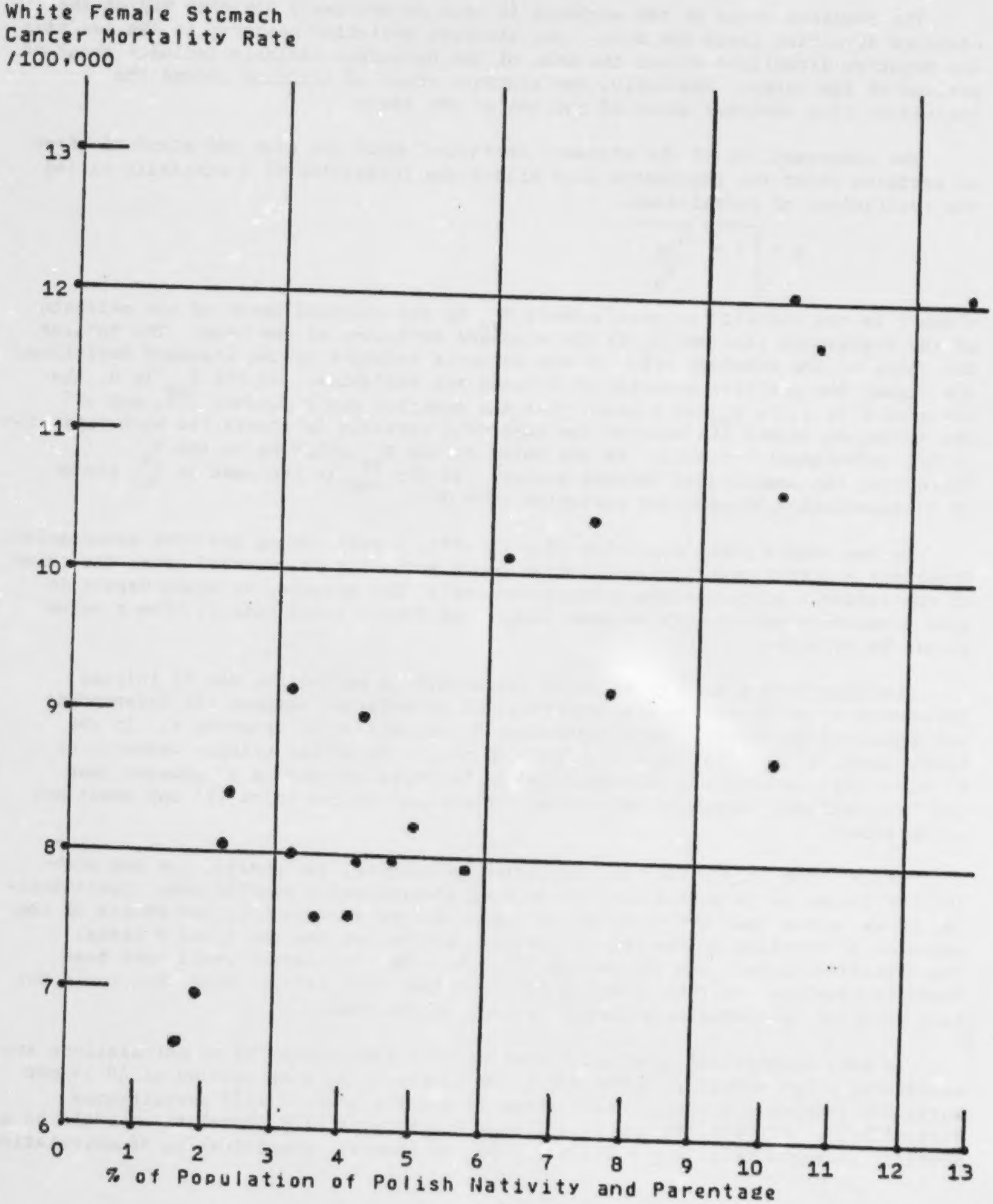
$$Y_e\ \text{Hudson} = 6.506 + 41.630\ (.10384) = 10.829$$

While the fit of the line to the data is pretty close, scatter about the line leads to an imperfect fit. The standard error of the estimate is used to measure the accuracy of the equation.

$$S_{ee} = \sqrt{\frac{(Y_a - Y_e)^2}{n}}$$

ASSOCIATION OF FEMALE WHITE STOMACH CANCER RATES AND POLISH FOREIGN BORN POPULATION IN NEW JERSEY, 1950-69

Figure 1-1



where S_{ee} is the standard error of the estimate; Y_a are the actual values of the dependent variable (white female stomach cancer); Y_e are the estimated values of the dependent variable; n is the number of cases (21, for example).

The standard error of the estimate is used in precisely the same way as the standard deviation about the mean. One standard deviation measured in the positive and negative directions around the mean of the dependent variable includes about 68 percent of the cases. Similarly, one standard error of estimate around the regression line includes about 68 percent of the cases.

The comparability of the standard deviation about the mean and standard error of estimate about the regression line allows the derivation of a statistic called the coefficient of correlation.

$$r = \sqrt{1 - \frac{S_{ee}}{S_m}}$$

where r is the correlation coefficient; S_{ee} is the standard error of the estimate of the regression line and S_m is the standard deviation of the mean. The smaller the value of the standard error of the estimate relative to the standard deviation, the higher the positive association between the variables. If the S_{ee} is 0, the value of r is 1. An S_{ee} of 0 means that the equation was a perfect fit, and all the variation around the mean of the dependent variable is associated with variation in the independent variable. As the value of the S_{ee} relative to the S_m increases, the association becomes weaker. If the S_{ee} is the same as S_m, there is no association between the variables ($r = 0$).

In the sample case, the value of r is .851, a very strong positive association. Sometimes a strong negative association among variables is expected (when the value of one variable increases the other decreases). For example, we would expect to find a negative association between income and female liver cancer. The r value should be negative.

How high does r have to be to be important? A subjective way of judging importance is to determine the percentage of covariation between the independent and dependent variables. This percentage is calculated by squaring r. In the sample case, r^2 is .724 (.851 x .851). Normally, in social science research an r^2 value that exceeds .25 is considered to be important and an r^2 greater than .50 is considered extremely important. However, r^2 values below .25 may sometimes be important.

Rather than rely solely on subjective judgements, the analyst can use probability theory to judge the significance of a correlation coefficient. Specifically, if we assume that the study areas' counties are a representative sample of the universe of counties in the United States, then we can use the t and F tests. These statistics tell you the probability that the correlation could have been randomly created. In this study we have retained only correlations that could not have occurred by chance more than 5 percent of the time.

In many statistical studies, including this one, thousands of correlations are calculated. For example, in Chapter 2 the geographical distribution of 48 cancer mortality records are correlated. These 48 records produce 1128 correlations [(48)(47)/2]. Probability theory tells us how many of 1128 correlations could be a specific strength (e.g., $r \geq = 0.25$; ≥ 0.50) by chance. Specifically, 56 correlations of at least 0.25 (significant at the .05 level) should be expected among the 1128 correlations calculated for the 45 area region. Half of the 56 (28) should be expected to be positive and half (28) negative.

Knowing how many correlations could be due to chance is important when studying the general pattern of the correlation results. In the same case, if the analyst found that 70 of the 1128 correlations were at least 0.25 and that half are negative, one would conclude that, in general, the 48 mortality records were not strongly associated. Some of the 70 significant correlations might be extremely imporant, but generally the diseases do not strongly covary geographically. On the other hand, if 300 or more correlations were found to be at least 0.25 and most of them were positively associated, then the analyst can tentatively conclude that most of the correlations did not occur by chance. A large number of significant correlations does not guarantee an explanation of the significant correlations. A strong mathematical association implies a search for an explanation, if one had not already been hypothesized to exist, it does not promise an explanation.

The higher the level of significance chosen, the fewer the number of significant correlations should be expected. In the above example, a .01 significance level should produce only about 11 significant correlations by chance. And an .001 significance threshold (equivalent to an r value of about 0.5 or greater in the sample case) should produce only 1 significant correlation by chance. These observations will be placed in a context in Chapters 2-4.

Summarizing, this study uses linear correlation analysis to measure the association between cancer mortality and high risk indicators. The linear correlation statistic is based on the assumption that one variable changes at a constant ratio to the other. The association can be direct (both increase) or inverse (one increases and the other decreases). The laws of probability can be used to help interpret the strength of correlations.

Dependent variables are usually expected to associate with more than one independent variable. In the case of white female stomach cancer, we assumed that industry, air pollution, and other high risk factors might add to the explanatory power contributed by the Polish population factor. Multiple regression and multiple correlation are used to incorporate more than one variable into the analysis. In the sample case, the following two variable, mathematical equation was produced.

$$Y_e = 6.661 + 0.018X_1 + 27.085X_2$$

where X_1 is the total employed in manufacturing in 1000's and X_2 is the proportion of the population that is Polish. In comparison to the single variable equation, this equation improves r from .851 to .907, a significant improvement.

When multiple regression and correlation analyses are made, it is important to try to distinguish between the importance of each of the independent variables. The relative importance of each variable is measured by the partial correlations. If the independent variables are themselves correlated, the partial correlation coefficients may be misleading. Intercorrelated independent variables present the analyst with a serious problem because equations with highly correlated independent variables are likely to lead to erroneous estimates.

When the independent variable data are highly intercorrelated, factor analysis may be used to cluster the variables and help overcome the above problem. Factor analysis is a mathematical procedure that restructures raw data into distinctive clusters. Factor analysis forms new variables that express the geographical coincidence of different attributes. For example, in the study area, counties that lie in the core of the urban-industrial corridor have a good deal of manufacturing, electricity generation through fossil fuels, automobile traffic, air emissions, a large proportion foreign born and unmarried females, and a water supply that generally comes from public potable surface water. The statistically constructed core variable which is called a "factor", is identified by two tables. The first presents correlations (called "factor loadings") between the core factor and variables:

Variables	Factor loading (r between factor and variables)
1. Population density	.928
2. Air pollution indicator	.921
3. Motor vehicle traffic indicator	.832
4. Fossil fuel burning	.775
5. Manufacturing indicator	.770
6. % Russian Foreign Stock population	.630
7. % Female married	-.541
8. % Polish Foreign Stock population	.437

The second table identifies those counties strongly manifesting a high concentration of industry, electricity generating stations, Jewish and unmarried female populations. A statistically standardized score was calculated for each of the 45 regions in the study area. Rather than display the list at this time, the ranks of the four counties with the highest values are indicated below: The four all lie in the study area's urban-industrial corridor.

Rank	County(ies)
1	New York City
2	Philadelphia
3	Hudson
4	Essex

Factor analysis is an important tool in this study. It more closely captures the geographical coincidence of some of the most important cancer mortalities and high risk factors than any single indicator. In short, it comes as close as we can get to mathematically approximating the real-world interaction of high risk factors.

Summarizing, this study uses correlation, regression and factor analysis to study the association between cancer mortality and high and low risk factors in a region centered on New Jersey but also including nearby portions of New York, Connecticut, Pennsylvania, and Delaware. The advantage of these methods are that they express associations in precise terms. The disadvantage is that the methods assume that extremely complicated relationships can be expressed by mathematical statistics.

False Positive, False Negative, and Other Problems

Correlation and regression analysis may produce false positive, false negative and confusing results. A false positive result occurs when a positive association is found between two variables that should be uncorrelated or should be negatively associated. An example of a false positive result which the author read is an equation which shows a positive association between health and atmospheric arsenic. It seems inconceivable that arsenic in the air improves health. A false negative result is finding a negative association when no asssociation or a positive association is expected. A false negative result would be that smoking habits and lung cancer are not positively associated. A confusing result is when one variable, among a group of highly intercorrelated variables, is associated as expected, but the other variables do not follow the assumed association. An example of a confusing result is that air emissions and disease rates are associated, but that motor vehicle traffic and industrial concentration are not related to the same disease. This section will explain why these problems occur and how they may be avoided or at least kept to a minimum.

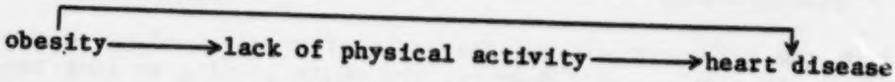
False positives and negatives and confusing results occur because of a poor choice of variables, poor quality data, limitations of the methods, and analyst errors. Sometimes the seemingly false results are not false; but rather the generally expected relationship does not hold for the specific study area under scrutiny.

False negatives and positives and confusing results are a serious problem in nearly every correlation and regression study. When correlation and regression analysis are involved, loud disagreements at professional meetings and lengthy letters to the editor are common.

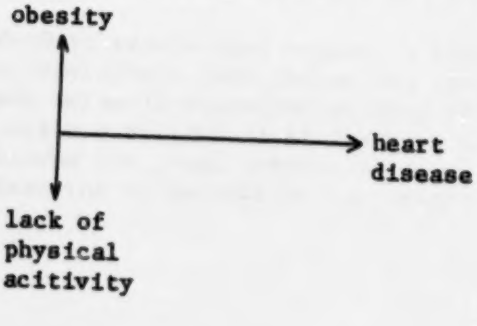
While data and methodological questions abound, the usual reason for the arguments is a lack of agreement about the direction of cause-and-effect. For example, what are the associations among obesity, physical activity and heart disease? One model implies that physical activity is the key causal variable, that is if you exercise, you greatly reduce your chance of heart disease.

lack of physical activity ——→ obesity ——→ heart disease

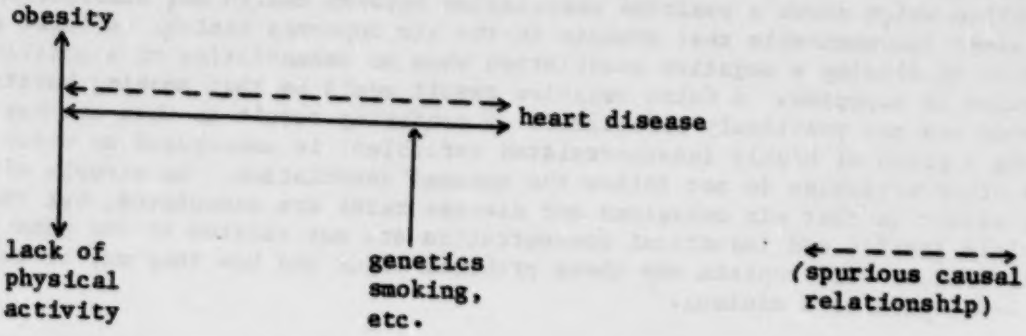
A second model implies that obesity in and of by itself leads to heart disease. In addition, obesity leads to a lack of physical activity, which in turn, increases the risk of heart disease.



A third heart model is that feedback occurs between obesity and a lack of exercise. Both lead to heart disease.



Finally, other analysts may argue that while obesity and exercise seem crucial, they are really manifestations of such real causal variables as heredity, smoking habits, quality of the home life, and other factors. According to this fourth model, the only reason obesity and a lack of physical activity correlate with heart disease is because the real causes of heart disease have been omitted from the model.



Perhaps even more so than for heart disease, the chain of causality leading to contracting cancer is extraordinarily complicated. Freeman, Levine, and Reeder (1972) suggest the following chain of events to explain cancer of the scrotum.

1. Nationality, social class, race, and other status : English lower class, white, urban
2. Occupation : chimney sweepers
3. Behavior pattern : working in intimate contact with soot
4. vehicle : soot, product of incomplete combustion
5. agent : unknown

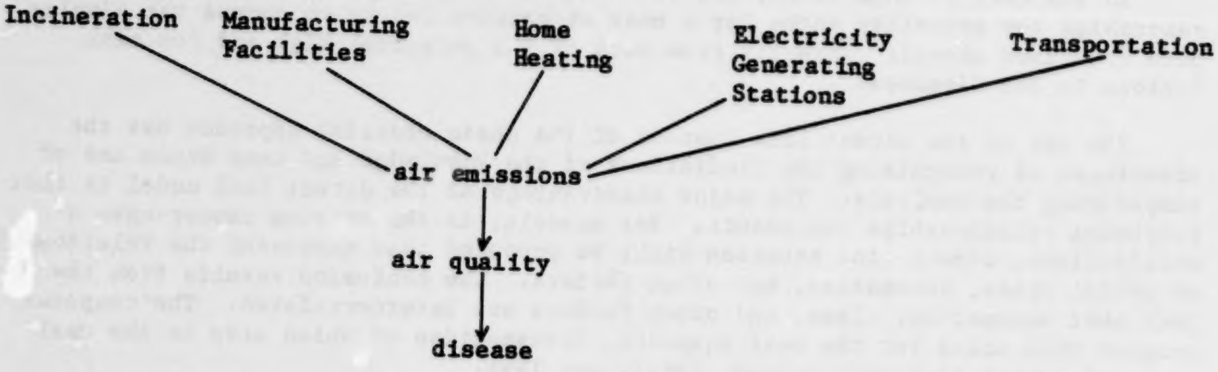
Susceptible host

7. tissue change : cancer of scrotum

Viewed from the 1977 perspective, the chain of events leading to cancer of the scrotum is relatively simple. For example, the chain leading to lung cancer among male whites in the United States would include not only nationality, occupation, behavior pattern, vehicle, agent, and susceptible host, but also include smoking habits, air quality, length of residence in the area, and other factors.

The author has adopted the complex view of cancer represented by Graham and Reeder. Adopting a perspective is, however, far easier than operationalizing the perspective. Two general approaches may be used to operationalize the chain leading to the development of cancer. One approach is to develop equations linking each of the steps in the chain. In the cancer of scrotum case, six equations would be required. The first would define how occupation is related to nationality, social class, race, and other status factors. Upon sucessful derivation of the first equation, a second equation would mathematically state the relationship between occupation and the behavior pattern that leads to the contraction of cancer. Once predictions of the behavior pattern could be made, then an equation would be derived linking the behavior pattern to the vehicle (in this case, soot). Next, an equation would link the vehicle to the agent and finally to cancer of the scrotum.

Deriving a chain of causation is difficult. If even one data set is missing or of poor quality, the chain cannot be completed. Sometimes, a step or two can be skipped without breaking the logic of the chain. More often however, breaking the chain leads to unacceptable results. For example, let us assume that an unnamed disease is related solely to air pollution. We might construct the following representation of this very simple chain in urban areas:

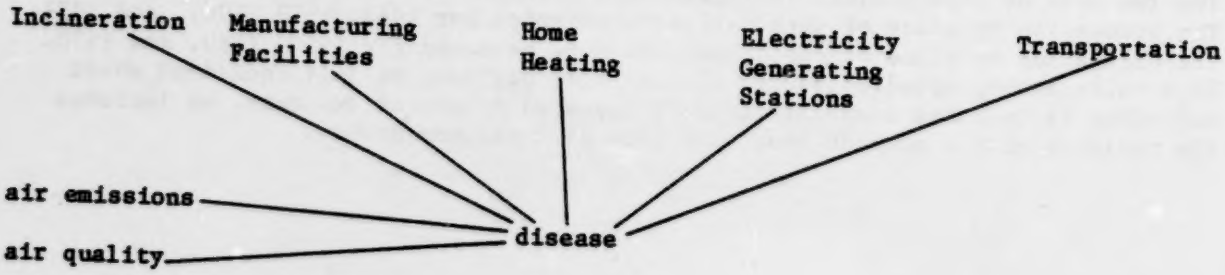

Incineration
Manufacturing Facilities
Home Heating
Electricity Generating Stations
Transportation
air emissions
air quality
disease

Let us further assume that the incineration, transportation, manufacturing, and disease data are good sets, but that the remaining data (home heating, electric generating stations burning fossil fuel, air emissions and air quality) are weak in spots.

Three equations would express the chain. The first is multivariate and ties air emissions to the five main sources of air pollutants. The second uses air pollutants and an air quality model to tie emissions to quality. The third equation ties air quality directly to the disease.

The problem with the three equation model is that correlating some good quality data with poor quality emissions and air quality data will probably produce questionable results which will ultimately be reflected in poor equations through all phases of the chain.

Given a mixture of poor and good quality data, the following simpler model is recommended.


Incineration
Manufacturing Facilities
Home Heating
Electricity Generating Stations
Transportation
air emissions
air quality
disease

On the one hand, the simple model does not explore the logical chain leading to the disease; on the other hand, it does not mix poor and high quality data through a chain.

Usually the chain is not as simple as the one pictured for cancer of the scrotum or for the unnamed disease. For example, whereas nationality, race, and social class often imply a particular occupation, a particular occupation sometimes leads to socioeconomic status. In short, the ethnicity/socioeconomic status link to occupation works in two ways, not just in one way. Accordingly, the equations should not lead directly from one to the other, but should contain feedbacks. Feedbacks are possible to build, but require more confidence in the data than we have in the data available to us.

In the case of this study, the uneven quality of the data and difficulty of describing the causality paths for a host of cancers led us to choose the simpler path (outlined above): directly from each of the potential high and low risk factors to the diseases.

The use of the direct link instead of the chain modeling approach has the advantages of recognizing the limitations of the knowledge and data bases and of simplifying the analysis. The major disadvantage of the direct link model is that confusing relationships can result. For example, in the scrotum cancer case a multivariate, direct link equation might be produced that expressed the relationship of social class, occupation, and other factors. The confusion results from the fact that occupation, class, and other factors are intercorrelated. The computer program will solve for the best equation, irrespective of which step in the real chain of events is first, second, third, and last.

If the analyst is cognizant of the high probability of finding confusing false results, preparations can be made to avoid or at least minimize the problems. The best solution to false positive and false negative problems and a solution that will help avoid bias problems is to adhere as closely as possible to the literature, but drawing causal hypotheses strictly from the literature can stifle creative research. If the researcher thinks that the literature ignores some factors, these should be included with accompanying arguments for and against the hypotheses. Otherwise, one should stick to the literature to avoid introducing false variables. The second best means of avoiding false and biased results is to make sure that the results consistently appear. The reader will have greater confidence in relationships that appear in three or more time periods than in a relationship that appears only once. Similarly, if the expected association appears in three geographic regions, as opposed to only one, the reader can have greater confidence that the relationship is not spurious.

This study has built-in time and space consistency tests, whenever possible. The two sets of occupational variables were gathered for multiple time periods. The occupation by place of work data were recorded for 1947, 1958, 1967, and 1972. The occupation by place of residence data were gathered for 1950, 1960, and 1970. If a relationship appeared during all the time periods, we felt confident about including it in later runs. If it only appeared in one or two runs, we included the variable with a comment about its lack of time consistency.

In general, all of the personal environmental data sets were collected for 1950, 1960, and 1970. The occupational data were available for 1947, 1958, 1967, and 1972 or 1950, 1960, and 1970. The historical record of local ambient environmental data varies. Data on fossil fuel combustion is the best set, covering more than thirty years. Motor vehicle travel data are consistent in New Jersey for thirty years, but spotty elsewhere. Air quality and air emissions data are sparse, and water supply data are available for only one time period.

Demonstrating consistency across different geographic regions is important to ensure that a small number of places are not dominating and biasing the relationship. For example, homoscedasticity is a mathematical problem diagnosed by the appearance of regression residuals that are consistently associated with a specific set of values of the independent variables. Regression residuals are the difference between the actual values of a characteristic of a set of observations and the values predicted by an equation. Assume that the actual median income values of two counties are $15,000 and $12,000. An equation using education, income, and age predicts the values to be $14,000 and $10,000. The regression residuals are $1,000 ($15,000 - $14,000) and $2,000 (($12,000 - $10,000). Studying these residuals may help you diagnose an unexplained variable(s) that should be added to the equation. On the other hand, it may tell you that some of the observations are so unique that they belong to a different population of counties and should be considered separately. In our case, for example, assume that a disease was correlated with the presence of industry. Yet, if the regression residuals for all the areas with a good deal of employment were high positive values, the analyst should suspect homoscedasticity. Perhaps, the high employment counties are a different group from the low employment counties.

Three methods are available for assuring consistency across geographic regions. One is to perform the analysis on the study area as a whole, and then repeat the analysis on a randomly drawn or spot samples of places in the study area. The second method is to make an educated guess about which places are likely to cause problems, and repeat the analysis without these places. The third method is to check the regression residuals for clustering of results by propinquity, size and other factors.

This study used all three methods. Initially, we made all our correlation and regression runs with the full 45 places. Next, a number of runs were made with 30 randomly chosen places from among the 45. Third, we expected that New York City and Philadelphia might dominate the results because of their large size measured in employment and emissions. Likewise, the presence of some counties with few citizens might produce results unduely influenced by the relatively few citizens in those counties. Accordingly, all runs that survived the first set of tests (with 45 cases) were rerun with 40, 35, and 25 of the 45 places. The most populous places (New York City, Philadelphia) and the least populated (Carbon, Monroe, and Pike) were excluded from the 40 area runs. The 10 least populous counties were excluded from the 35 case runs, the 20 least populous from the 25 case runs. Finally, since New Jersey is the center of the study area and is the major are of interest of the study's sponsors, the runs were repeated with the 21 New Jersey counties. Overall, by repeating the correlation runs for randomly drawn places and spot samples and by checking residuals, we tried to assure the geographic consistency of the results.

If a lack of consistency is found, a variety of countermeasures are available. Sometimes the data may be transformed to logarithms to counteract the dominance of a few places. Second, runs may be made without the dominant places included in order to observe the relationship without their presence. And sometimes, a lack of consistency is a sign that important factors have been omitted. If they are included, the problem may disappear.

The results should not only consistently appear, but they should also be consistently strong. "Strong" is a subjective term that can be quantified in statistical studies. One way of quantifying "strong" is by assuming that the study area is a representative sample of the universe of the counties in the United States. If this assumption is made, we can choose to set minimum correlation values that must be met for the variable to be considered to be significant. In this study we chose to make the assumption that an .05 level of significance was strong enough for continuation of a variable in the analysis. Even if the association between a cancer variable and a high risk factor was in the expected direction, if the correlation could have occured by chance more than one out of every twenty times, we eliminated the variable from further consideration at the bivariate correlation stage. At the multivariate analysis stage, variables were not allowed into the equation unless they produced an F value $\geq$ 1.0.

Next, some high risk indicators were eliminated at the bivariate screening stage because, while they produced some expected significant results, they also produced some unexpected and obviously spurious results. For example, the indicator percent employed in construction by place of residence produced a significant positive correlation with white female lung and skin cancer. However, it also produced results which suggested that some cancer rates would be decreased by increasing the role of the construction industry in the region's economy. Perhaps, we have lost some valuable information by deleting the variable. On the other hand, we have avoided potential false positives. Fortunately, few of these variable deletions were necessary.

The strength of an association among dependent and independent variables can be judged not only against a level of probable random occurence, but also against the strength of other correlations. Some associations should be stronger than others. For example, we would expect that the number of people employed in an industry would associate more strongly with cancer than would the number of establishments. Similarly, we would assume that the number of production workers in an industry would correlate higher than the total labor force which includes office personnel.

We would expect that the presence of Jewish people would be associated with elevated stomach, intestine, pancreas and some other cancer rates. However, we expect the strongest association of Jewish people should be leukemia. Similarly we assume that the highest association between the presence of Polish-born populations and cancer to be with stomach cancer.

The researcher can hypothesize a series of expected results for control factors which, if realized, strengthen the case. The controls can assume a variety of forms. Two of the most important controls used in this study will be briefly explained: fallback and control diseases. We call the first the "fallback to the general category" guideline. It was asumed that the presence of industry would correlate with elevated cancer rates. But industry is spatially agglomerated, and therefore it might not be possible to separate the effects of different industries. The guideline is to use a general category unless the more specific category consistently demonstrates higher correlations than the general category. An example will illustrate the application of the guideline. Let us assume that correlation analysis produced the following hypothetical results:

Correlation of Industry and Lung Cancer

Independent Industry Variables	Year 1947	1958	1967	1972
Textiles	.480	.430	.410	.420
Metals	.520	.500	.530	.430
Petroleum	.495	.460	.385	.435
Chemical	.540	.450	.415	.430
All Industry	.500	.450	.425	.425

All industry is the general category; textiles, metals, petroleum and chemicals are the specific categories. All the correlations are statistically significant at the .05 level for 45 cases and all should be expected from the literature. Therefore we have no reason to reject any of them. Yet, we cannot include all of them as independent variables without producing confusion because the independent variables are likely to be highly intercorrelated. Examining the bivariate correlations, one finds that only the metals industries produce higher correlation values than the all industry category. Therefore, textiles, petroleum, and chemical would be deleted after the bivariate stage and all further analyses would use metals and all industry as working variables.

The implication of following the fallback guideline is that some false positives will be avoided at the cost of causing some false negatives. In the above example, there is strong evidence to suggest that the manufacture of certain petrochemicals and other chemicals cause lung cancer. While the high bivariate correlation between the presence of the petroleum and chemical industries and lung cancer would be pointed out, they would be excluded from the later multivariate analysis. The fallback approach is not the only control procedure we used. It is, however, an important guideline based on the relative strength of correlations that will help avoid false positives.

A second important means of controlling for false positives and negatives is the introduction of diseases which according to the literature should not be correlated with the high risk factors. If a strong relationship is found between the diseases of interest and the high risk factors, while no correlations are found between the control diseases and the high risk factors, then one has additional evidence to argue that the strong results are not spurious. The use of control diseases has one disadvantage. If the control disease unexpectedly correlates with the high risk factors, the analyst has false positive results to explain.

Only two control diseases were available for this study: prostate and female breast cancer. We expected no significant correlations between prostate cancer and the high risk factors. We expected positive correlations between female breast cancer and high socioeconomic status, but no associations between female breast cancer and occupational and local ambient environmental exposures. Sometimes, irrespective of the care one takes in screening variables through literature and statistical phases, apparently false or at least unexplained associations are passed through to the final steps. When this problem occurs, this study has taken two steps. First, the results will be calculated with the supposedly spurious variables included. Then, the results will be calculated without the unexpected variables. All the important combinations are reported.

False negatives are as formidable a problem as false positives, and they may even be more difficult to control than false positives. With respect to correlation analysis, false negatives normally arise because of three reasons: (1) a warranted negative finding; (2) poor quality data; and (3) the nature of correlation analysis. An unexpected negative finding may be justified. A relationship between the presence of known carcinogens in the environment and cancer does not always occur. For example, while asbestos is a proven carcinogen, there is yet no evidence to suggest that cancer mortality rates are higher than average in places that have higher than average levels of asbestos in their water (e.g., Duluth) or in the air (e.g., Rockville, Maryland).

Often, however, false negative and false positives are produced because of the use of poor quality data. Two examples drawn from this study will illustrate the problem and the solutions to it. The air quality, air emissions, and water supply data are the weakest information sets that were used in our statistical analysis. The data could not be tested for temporal consistency and some of the information had to be estimated. Some compensation was, however, possible. Specifically, one begins by assuming that the locations of emitters (industry, fossil fuel generating stations, and motor vehicles) and air emissions should be correlated. Next, air emissions shuld be associated with air quality. If these two assumptions are valid, then correlations among cancer and the presence of the emitters may be used to shore-up weak air quality and emissions data In our case study, the fossil fuel and industry data support the air quality and emissions data.

Sometimes when no causally linked data are available, the analyst can turn to control diseases to try to avoid a false negative or positive. For example, if smoking data are not available, emphysema may be used as a surrogate for smoking. If emphysema data are not available, then lung cancer may act as a surrogate for smoking when bladder, stomach, and other smoking related cancers are studied. However, in general, and particularly in the sample case, it is troublesome to use one disease (e.g., lung cancer) as a surrogate for one environmental factor (e.g., smoking), while elsewhere in the study you are testing for the association between the control disease (e.g., lung cancer) and other variables (e.g., non-smoking).

Even the best data can produce unexpected and unwarranted negative results because of characteristics of correlation and regression analysis methods. Correlation and regression analyses and many other statistical techniques are insensitive to high risk factors that must be present to cause the disease but which do not vary among regions. Probably the best illustration that we might encounter is among cigarette smoking, air pollution, and lung cancer. We have no cigarette smoking data in the Part I report. Therefore, this problem has not arisen in this study. However, since we hope to develop cigarette smoking data for Part II, it is worthwhile to explore what might happen. Two sets of hypothetical data are presented below.

Data Set A

	Lung Cancer Mortality Rate/100,000 (1)	Air Quality, TSP ug/l (2)	Per Capita Cigarettes Smoked (3)
County 1	50	40	2
County 2	60	45	4
County 3	70	75	5
County 4	80	100	8
•			
•			
•			
•			
•			
•			
County 100	60	55	4

Data Set B

	Lung Cancer Mortality Rate/100,000 (1)	Air Quality, TSP ug/1 (2)	Per Capita Cigarettes Smoked (3)
County 1	50	40	3
County 2	60	45	3
County 3	70	75	3
County 4	80	100	4
•			
•			
•			
•			
•			
•			
County 100	60	55	3

Data set A is the expected. Counties with high rates of lung cancer have poor air quality which enhances the effects of cigarette consumption. Data set B portrays a very different picture. Lung cancer and air quality seem to be correlated; but per capita cigarettes smoked is almost a constant and does not correlate with lung cancer. Cigarette smoking and lung cancer are not correlated in data set B. Their association, however, may be crucial. Exposure to cigarette smoking may be a necessary ingredient of lung cancer; but because smoking does not geographically vary in the illustrative data set, it will not be correlated with lung cancer. In short, the cocarcinogenic effects of agents will not necessarily be isolated by correlation and regression techniques, nor by many other statistical methods.

Even if the analyst has controlled against false positives and negatives, the use of intercorrelated independent variables can cause confusing false negatives and positives to occur in regression equations. If the user forces highly intercorrelated variables into an equation, the mathematical process will change the sign of one or more of the variables in order to accommodate the requirement that the resultant equation be the best mathematical fit.

Three general approaches are used to control this so-called "multicollinearity" problem. One approach is to use one of the highly intercorrelated variables as a surrogate for the set of correlated variables. The other risk factors can be prevented from entering by controlling the F entry value. The F statistic is a means of allowing only relatively uncorrelated independent variables into the equation by demanding that each new indepedent variable explain a minimum amount of unexplained variation in the dependent variable. For example, air emissions, motor vehicle traffic, generating stations, and industrial facilities are likely to be highly intercorrelated. One may have to be used as a surrogate for the others. Normally, the most general variable should be used as the surrogate.

A control variable may be used to assist in defining the best surrogates. In this study, we have used population density as the control variable. Population density, in and of itself, should not be a factor contributing to high cancer risk, unless indirectly by contributing to stress it enhances the chances of contracting cancer. However, as a surrogate for urban-industrial development, it should be highly intercorrelated with the high risk occupational and local ambient environmental factors that are involved in carcinogenesis. Accordingly, we have used population density as a measuring stick against which we compare other indicators of urban-industrial development (e.g., industry, generating stations, highways, and others). If a variable, for example, the presence of industry, consistently produces a higher correlation with the disease of interest than does population density, we assume that it is a good general surrogate and also has some unique contribution to make. Variables that produce lower correlations than population density may have a unique contribution to make but are not considered to be good general surrogates.

A second approach used to avoid confusing regression models is to employ factor analysis to define surrogate variables for highly intercorrelated high risk factors. The statistically constructed factors can be made to be statistically independent.

The third approach that has been used is the calculation of average values of independent variables when the data are available for many time periods. For example, let us assume that the following significant bivariate correlations are produced.

Variables	Correlation Value
Cancer rate and manufacturing employment:	
1947	.600
1958	.570
1967	.550
1972	.560

Should all four be used? Or should 1947 be used because it is the highest? Or should 1967 be used because it is the lowest? If all four variables are inserted simultaneously into the model, the stepwise statistical process because it selects only relatively uncorrelated independent variables will initially accept the highest and probably not the others, unless the analyst forces them in. The result may be a tendency to overestimate the strength of the relationship between manufacturing and cancer. In order to avoid this overstatement problem, we have computed the averages of variables for which we have more than one time period of record. In the manufacturing case above, we would average the manufacturing employment in 1947, 1958, 1967, and 1972 and use the resultant average datum to represent manufacturing employment for the entire time period.

After controlling for false negatives and positives, the analyst can further test the results by substituting values of the independent variables into the equations. If unexpected major changes in the cancer rates result, the equations are suspect. While we have made some equation tests, we have great reluctance to use the equations for anything other than analysis of residuals for two reasons. They imply specific relationships between cancer and the independent variables that are difficult to justify given the poor quality data that sometimes had to be used. Second, one independent variable may be operating as a surrogate for others.

An example should illustrate why we made this important decision. Suppose the following hypothetical equation was derived:

$$L_{cm} = 20 + 3\ \mathrm{MVMT}$$

where L_{cm} is male lung cancer mortality and MVMT is million vehicle miles travelled.

County X had 8 million MVMT. Substituting, we derive

$$L_{cm} = 20 + 3(8)$$
$$= 48$$

The derived number, 48, may be a reasonable estimate of the actual lung cancer mortality among males in the county. However, forecasting a change in lung cancer mortality on the basis of this equation would be bad science for two reasons.

First, on the one hand, motor vehicle miles may be increasing and therefore a higher lung cancer rate should be expected. On the other hand, automobile emissions have become progressively less carcinogenic, if measured by benzo(a)pyrene and other hydrocarbon emissions. Accordingly, an increase in motor vehicle travel may not equate with an increase in auto emissions.

The second reason for refraining from the use of equations for forecasting is that as indicated above, motor vehicle miles driven may be a surrogate for a host of urban-industrial variables. In addition to what happens to motor vehicle travel, the analyst would have to incorporate changes in industrial location, fossil fuel generating stations, smoking habits, and other factors to make a reasonable forecast of changes in the cancer rate. Overall, our viewpoint is that the great complexity of carcinogenesis, when combined with data limitations, precludes the use of the forecasting equations produced by the study.

The final pitfall to avoid is drawing conclusions about individuals on the basis of data derived at the community, county, regional, and state scales. The further the analysis is removed from the individual case, the more likely the ecological fallacy of mistakenly attributing specific relationships to individuals based on analysis of general populations is likely to occur. An example will be used to illustrate this potential problem. Let us suppose the following correlation values are produced at the county scale

Lung cancer mortality and metal workers	.500
Lung cancer mortality and Russian born population	.450

Can we conclude that both metal workers and Russian born population are contracting cancer? Can we conclude that Russian born metal workers are the population at risk? We probably can conclude all of the above, but not necessarily. If the literature identifies metal workers and Russian born people as high risk groups, then the analyst can have some confidence in the results. But one cannot be positive, unless death certificates are examined to determine if in fact people who died from lung cancer were metal workers and were Russian born. Death certificates are not difficult to obtain but require substantial funds to carefully examine. Moreover, cause of death diagnoses are imperfect for many diseases (e.g., stomach, pancreas, prostate), and other of the information contained on certificates is often inaccurate. In short, the results that we produce are suggestive of hypotheses but are not epidemiologically based conclusions.

Summarizing, this section has reviewed the statistical tools we employed, the pitfalls of using the tools, and methods of controlling or at least minimizing the pitfalls. The next section presents the data that were statistically analyzed.

A Description of the Data

Cancer Mortality Rates

The National Cancer Institute reviewed the public use tapes for 1950-1969 in order to compute age-standardized rates in 5-year cohorts by race and sex for all United States counties. They then described the spatial variation of 35 cancer body sites (Table 1) by county for over 3,000 counties in the 48 contiguous states. The NCI was forced to resolve three problems: accuracy of diagnosis; revisions of the international classification of diseases in 1958 and 1968; and population growth.

United States vital statistics data are gathered from individual death certificates. A considerable amount of information recorded on the certificate is not available. The National Cancer Institute was able to obtain nine items of information from computerized summary tapes of the death certificates developed by the National Center for Health Statistics. The tapes include the following data:

1. State and county occurrence of death;
2. State, county and city of residence;
3. Age, sex, race; and
4. Underlying cause of death

The availability of this information on computer tape offered the NCI researchers the opportunity to prepare cross-tabulations of cause by place, sex, race or age. The NCI was able to obtain tapes back to 1950; tapes available to the public extend back only to 1958.

The objective was to compute twenty yearly-rates for all the counties in the 48 contiguous states. The difficulty in computing rates is that two sources of data are used, one the Vital Statistics, and two the Census. In order to obtain yearly intercensal estimates of county population, the method chosen was to compute a year by year linear interpolation between the two census years.

In order to remove the effects of age, sex and race, separate rates were calculated. Then, the tape data, which are in yearly cohorts, were aggregated into five-year cohorts. While there is some uncertainty as to the amount of information lost by grouping into 5-year cohorts, Barclay (1958) suggests that the effects of the age structure do not significantly bias such aggregations.

Because cancer mortality data were available by 5-year age-cohorts for each place, a direct method of age-standardization was used based on the 1960 total population. The formulas used by the NCI are presented below (Mason and McKay, 1974):

$$S(i,t) = \frac{D(i,t)}{R(i,t)} \cdot 100{,}000$$

$$A(t) = \frac{\sum_{i=1}^{n} S(i,t)\,P(i) \cdot 100{,}000}{\sum_{i=1}^{n} P(i)}$$

$S(i,t)$ = The age-specific death rate in age group i over t years
$A(t)$ = The age-adjusted death rate over t years
$D(i,t)$ = The number dead in age group i over t years
$R(i,t)$ = The number in age group i in the population at risk over t years
$P(i)$ = The number in age group i in the standard population in the base year
t = The number of years observed, which is 20
i = A particular age group m years wide
n = The number of age groups

TABLE 1

Cancer Sites Mapped and Tabulated at the County Scale by the National Cancer Institute 1950-1969 Study

Cancer Sites ICD	Description
140	Lips
142	Salivary Glands
146	Nasopharynx
141, 143, 144, 145, 148	Tongue; Floor of Mouth; other parts of mouth and mouth unspecified; oral mesopharynx; pharnyx, unspecified
150	Esophagus
151	Stomach
153	Large Intestine except rectum
154	Rectum
155	Biliary passages and liver (stated to be primary site)
157	Pancreas
160	Nose, nasal cavities, middle ear, and accessory sinuses
161	Larynx
162, 163	Trachea, and bronchus and lung specified as primary; and lung and bronchus, unspecified as to whether primary or secondary
170	Breast
171	Cervix uteri
172, 173, 174	Corpus uteri; other parts of uterus including chorionepithelioma; and uterus, unspecified
175	Ovary, fallopian tube and broad ligament
177	Prostate
178	Testis
180	Kidney
181	Bladder and other urinary organs
190	Melanoma of skin
191	Other skin
192	Eye
193	Brain and other parts of nervous system
194	Thyroid glands
195	Other endocrine glands
196	Bone (including jaw bone)
197	Connective tissue
201	Hodgkins's disease
200, 202, 205	Lymphosarcoma and reticulosarcoma; other forms of lymphoma (reticulosis); mycosis fungoides
203	Multiple Myeloma (Plasmocytoma)
204	Leukemia and Aleukemia
147, 152, 158 159, 164, 165, 176, 179, 198, 199	All ICD's not previously listed
ALL	All malignant neoplasms

For any given site and race, an individual county was included if there was at least one death for either sex. Blanks were used to denote no mortality for the particular site, rate and sex. If the county rate was less than 0.049 per 100,000, it was reported 0.0. All state rates have been reported to two decimal places.

All rates were converted to the sixth revision of the international classification of diseases. In a conversation with Frank McKay, one of the authors of the NCI study, he told us that some of the conversions were of dubious use since there had been major changes in several categories between the 6th and 7th revisions.

In New Jersey, deaths by cause by race were not collected for the years 1962 and 1963. For these years, the NCI study averaged a four-year period, two years prior to 1962 and two years after 1963, computed a rate of change and interpolated for 1962 and 1963. Not accounted for were possible incidences of undercounting in the census, particularly for the nonwhite populations. An undercounted population group could, in reality, be experiencing higher-than-reported mortality rates, or lower-than-reported rates could also represent the situation.

The NCI study offers a single, age, sex, and race standardized rate for a twenty year time period. We would have preferred standardized and unstandardized rates covering shorter time periods beginning in the early or mid-1960's. These are available for Parts II and III of the study. In addition, standardized and unstandardized control diseases would have proven useful. Specifically, the following mortality records would have helped us: influenza and pneumonia, emphysema, suicide, infant mortality, symptomatic, senile, and ill-defined. Influenza and pneumonia would have helped identify places where an unusual number of people might have died from diseases that claim people before they die from cancer. These two causes of death are frequently used to isolate annual fluctuations for other causes of death. Emphysema may be used as a control disease for cigarette smoking. Suicide and perhaps infant mortality should not be strongly associated with occupational and local ambient environmental causes. Finally, the combination of symptomatic, senility, and ill-defined suggests the care with which the diagnoses are made.

Overall, the dependent variables in the study are the NCI's 1950-1969 age, sex, and race standardized mortality rates by county. The fact that the cancer rates are for the period 1950-1969 led us to seek independent variables for the period 1930-1970. The variables that were gathered are reviewed in the next section. Each of the subsections has four general parts: (1) data sought, (2) data actually available and obtained; (3) variables developed; and (4) associations expected between cancer mortality and surrogates for high risk factors. As an aid, Table 2 summarizes the highest hypothesized risk factors and ties them to the data discussed in this chapter and to the body sites reviewed in chapters 2 and 4.

TABLE 2

Summary of Strongest High Risk Cancer Factors

Category	Body Site**	Comment
1. **Questionable Rates**		
a. Different age distribution	female breast; leukemia	May undermine analysis
b. Questionable diagnosis	pancreas; prostate; stomach?	May undermine analysis
2. **Host: Genetic and Prior Health**		
a. Family history	stomach; large intestine; lung; female breast; prostate?; leukemia	Polish?; German?; USSR?; Italian?
b. Metabolic abnormality	bladder	no data
c. Hypertension; Caucasian	female breast	no data
d. Diabetes	female breast; pancreas	no data
e. Blood group A	stomach; female breast; leukemia	USSR?; Polish?; German?
f. Familial polyposis of the colon	large intestine; rectum	no data
g. Immune globuin deficiencies; Gardner's syndrome Ulcerative colitis; Villous adenoma; Crohn's disease	large intestine	no data
h. Mongolism (Down's syndrome); Panconi's syndrome; Bloom's syndrome; Viral infection of mother	leukemia	no data

TABLE 2 (continued)

Summary of Strongest High Risk Cancer Factors

Category	Body Site**	Comment
i. Serum hepatitis; cirrhosis; hemochromatosis; parasitic infestations	liver	no data
j. Veneral disease?	prostate	no data
k. Pernicious anemia; achlorhydria; low pepsin levels	stomach	no data
l. Light skin color	skin	no data
3. Personal Environment		
a. Smoking, especially cigarettes	lung, stomach, kidney, liver, pancreas, bladder	no data*
b. Excess alcohol	liver	no data*
c. Drugs	lung, leukemia	no data
d. Excessive exposure to sunlight	skin	no data
e. Medical habits:		
1. radiation treatment	leukemia, skin, female breast	no data
2. low frequency of checkup	Excess or early mortality because of late detection	no data
f. Affluence:		
1. High	female breast, leukemia?	Socioeconomic status data: income, education, professional and managerial occupation
2. Low	lung, stomach, liver, bladder	

TABLE 2 (continued)

Summary of Strongest High Risk Cancer Factors

Category	Body Site**	Comment
g. Nutritional habits:		
1. High in fats	female breast, pancreas	no data
2. High in fats, refined carbohydrates, low in fibers, fresh fruits and vegetables	large intestine, rectum	no data
3. Malnutrition, especially of proteins; aflatoxins in foods, too much fat and Vitamin A consumption; food additives, pesticides in food	Liver	no data
4. Excess smoked fish and pickled vegetables; low in uncooked vegetables fresh meat, and Vitamins A and C	Stomach	no data
h. Poor living habits: smoking, poor dietary habits	All body sites	Unmarried males and females
i. Estrogen use; early age at menopause; no children, late age having children, few pregnancies	Female breast	Fertility ratio; persons per room
j. Ethnic and racial status:		
1. Nonwhite	Prostate	Standardized nonwhite rate
2. Polish	Stomach, lung	Polish
3. Jewish	Stomach, lung leukemia	USSR, German?
4. Occupational Environment		
a. Inorganic Chemicals:		
1. Arsenic	Lung, liver, skin	Employed in chemical, metals, agriculture, and leather tanning*

TABLE 2 (continued)

Summary of Strongest High Risk Cancer Factors

Category	Body Site**	Comment
2. Beryllium	Lung	Metal workers
3. Cadmium	Prostate?	Metal, nuclear plant* workers
4. Chromium	Lung and Other	Textile, chemical, metals and stone, clay, glass* workers
5. Iron	Lung and Other	Miners* and metal workers
6. Nickel	Lung and Other	Metal workers
b. Fibers and Dust:		
1. Asbestos	Lung, stomach, large intestine, rectum, and other	Miners*, construction, textile workers
2. Leather	Bladder and other	Textile and leather* workers
3. Talc	Lung	Miners*
c. Organic chemicals:		
1. Coal, soot, tar, and other coal combustion products	Lung, bladder, skin, and other	Electricity generators*, miners*, construction, petrochemical, and metal workers
2. Oils including petroleum, lubricating, paraffin, wax, anthracene, mineral oils and creosote	Lung, skin, and other	Textile, petrochemical workers
3. Chlorinated hydrocarbons (e.g., bis(chloromethyl)ether, chloromethyl ether; 4,4' -methylene-bis (2-chloroaniline); vinyl chloride chloroprene; auromine; trichloroethylene)	Lung, liver and other	Textile, chemical, farmers, leather* workers

TABLE 2 (continued)

Summary of Strongest High Risk Cancer Factors

Category	Body Site**	Comment
4. Others: benzene	Bladder, leukemia and other	gasoline stations*, food*, shoe*, chemical workers
4-aminobiphenyl	Bladder, liver	Chemical workers
2-(acetylamino) fluorene	Liver	Farmers, chemical workers
1-&-2 napthylamine, magenta, benzidine & salts, 4-nitrobiphenyl	Bladder	Printers*, textile, chemical workers
d. Physical agents:		
1. Ionizing radiation Radium, Radon, X-rays Uranium	Leukemia, lung, skin and other	Chemical workers, medical personnel*, miners
2. Nonionizing radiation Ultraviolet rays	Skin	Farmers, fishermen*, sailors*
5. Local Ambient Environment		
a. Chemical air pollution	Lung, stomach, liver? large intestine? pancreas?, kidney?, rectum?, bladder?, skin?, leukemia?	Air quality and air emissions; industry workers; motor vehicle travel; electricity generating stations
b. Physical air pollution from radioactive fallout	Leukemia	no data
c. Chemical water pollution and soil contamination	Stomach, large intestine, rectum, liver, pancreas, kidney, bladder, skin, leukemia	Water supply data

TABLE 2 (continued)

Summary of Strongest High Risk Cancer Factors

Category	Body Site**	Comment
d. Physical and biological water pollution from radioactive fallout and microbe water pollution	Leukemia and other	no data

---

*No data for this category, some is being developed for the Part II Study.

?Questionable relationship.

**Leukemia is a systemic disease, the others refer to specific body sites.

Personal Environment

Whereas the literature contains references to numerous personal factors that are assumed to increase an individual's chances of contracting cancer, surprisingly little data have been available to control for many of these factors in county scale studies. As previously indicated, we could not find cigarette smoking data. While we are planning to try to estimate cigarette smoking consumption patterns from survey data in the next phase of the study (Part II), we have none at this time. Not only were smoking data unavailable, but a control disease that should be associated with smoking (e.g., emphysema) could not be created for the 1950-1969 period.

Smoking is not the only missing data set. Alcohol consumption or a surrogate disease (cirrhosis of the liver) and drug data were not available. Some data on the use of X-ray equipment has been uncovered, however it will require a major effort to tabulate it. Nothing was available on the presence of pesticides, asbestos and other carcinogenic substances in food, in the home and in local recreation areas, on people's sunbathing habits, and directly on their eating habits. Finally, we have been able to develop a hospital availability data set, but have not yet been able to aggregate the data to the appropriate geographical scale.

Our personal environment data set for the first phase of the study is restricted to U.S. Census Bureau data. The census data allow us to test some hypotheses related to socioeconomic status and ethnicity.

High socioeconomic status is associated with high rates of breast cancer among white females, and low socioeconomic status should be correlated with high rates of lung, stomach, liver, and bladder cancer. The following eleven variables were collected for 1950, 1960, and 1970 as socioeconomic status indicators:

1. median income, family and unrelated individuals
2. median school years completed, males ($\geq$ 25 years old)
3. median school years completed, females ($\geq$ 25 years old)
4. % male high school and 1-3 years of college completed
5. % female high school and 1-3 years of college completed
6. % male at least college education
7. % female at least college education
8. % male, less than 4 years high school
9. % female, less than 4 years high school
10. % professional and technical employment of total employed
11. % managerial employment of total employed

Many of the socioeconomic variables are highly associated. Our intent was to find and use the few that fit the best.

The presence of four ethnic groups, in 1950, 1960 and 1970 was recorded using the 1950 Census "Country of Birth" heading the 1960 and 1970 "Country of Origin Headings." The ethnic population represents a combination of the first and second generations. The first generations were born outside the United States. The second generation are those at least one of whose parents was born outside the United States. In the case of mixed foreign parentage, the father's country of origin is used.

Four ethnic variables were created:

12. % Polish
13. % U.S.S.R. (Surrogate for Jewish)
14. % German (weaker surrogate for Jewish in parts of the study area)
15. % Italian

Ethnicity is primarily a surrogate for cultural factors including diet, smoking, drinking, medical habits, and many others. The following associations were expected:

12. % Polish	High positive association with cancer of the stomach, lung Low association with rectum, liver, pancreas, kidney, female leukemia Negative association with female breast, female bladder
13. % U.S.S.R.	High association with cancer of the stomach, lung, leukemia Low association with large intestine, rectum, pancreas, kidney Negative association with female breast, female bladder
14. % German	Low association with cancer of the stomach, leukemia Negative association with female bladder
15. % Italian	Low association with cancer of stomach, male large intestine, male rectum, male liver Negative association with female breast, female and male bladder

Among the four ethnic groups, we expected the strongest associations with the Polish population and weakest with the German population.

Finally, while the ethnic groups have been placed under the personal environment label, interaction was expected between the ethnic groups, occupational, and environmental factors. Specifically, preliminary working of the data mixed with some knowledge of history of the study area suggest that the Polish and Italian populations have settled in sections of the region that are degraded from the environmental perspective. In addition, these two populations have been more prone than the USSR and German populations to work in factory employment and to be less affluent. In short, if the ethnic associations hypothesized above are modified due to occupational and local ambient environment exposure, the expected correlations of the Polish and Italian populations should be increased, while those of the USSR and German populations should be decreased. For example, strong associations are expected between the presence of Polish and USSR populations and stomach and lung cancer. Due to interaction with occupational and local ambient environmental factors, the Polish relationships, especially among the male population, should be higher than those for the USSR population.

Three other sets of variables were included to test specific hypotheses suggested by the literature. Two variables were added to test the hypothesis that lengthy exposure rather than age leads to high cancer rates. The two variables were obtained for 1950, 1960, and 1970:

16. % same house as five years earlier; and
17. % different house, same county as five years earlier

We expected a positive correlation between the two variables and all cancer rates. However, since the cancer rates have been age standardized, the relationships should only be expected to be manifested among populations that are exceptionally stable.

Next, four variables were gathered to test the hypothesis that married persons are a lower risk group than single persons:

18. % male single among those more than 14 years old
19. % female single among those more than 14 years old
20. % male married among those more than 14 years old
21. % female married among those more than 14 years old

Two variables were used as possible approximations for the association between number of pregnancies among females and breast cancer.

22. Fertility ratio (only available for 1960 and 1970)
23. Persons per household, 1950, 1960, 1970

We expected that a high fertility ratio and large numbers of persons per household would be associated with low female breast cancer rates.

Summarizing, twenty-three variables were collected to test hypothesized associations between cancer mortality and personal environmental factors. Unfortunately, specific information about smoking, drinking, and other important habits could not be included among these variables due to the absence of the data. Attempts are being made to incorporate them into the second study.

Occupational Environment

Workers are exposed to the carcinogenic substances exemplified by those summarized in the occupational environment section of Table 2. Ideally, we would have liked to have had access to a historical record of specific chemicals used by facilities in the study area. Stanford Research Institute's (SRI) studies come closest, but are far from such an information bank. Accordingly, we fell back to the United States Census of Manufactures, County Business Patterns, and the United States Census of Population. The more than 700 variables we collected and derived reflect the great difficulty of trying to approximate specific exposures with general survey data.

We chose and derived twelve occupational categories that reflect the finest common level of detail available for the study area's 49 counties:

SIC Code	Description
15	All construction, including SIC codes 15, 16, and 17
22	Textile mill products
28	Chemicals and allied products
29	Petroleum and coal products
33	Primary metal products
34	Fabricated metal products
35	Machinery, except electrical
37	Transportation equipment
AGR	Agricultural workers
Sum of 3 most important codes	Sum of codes 28, 34, and 35
Sum of 7 industry codes	Sum of codes 15, 16, 17, 22, 28, 33, 34, 35, 37
Sum of all	All manufacturing

One of these codes, SIC 29, had to be dropped because of a lack of data. It will be included in the Part II study. The absence of historical data on petroleum and coal products forced us to rely on SIC Code 28, chemicals, as a surrogate for both industries 28 and 29. Presently, based on experience with water pollution permits, we believe that there is a good fit between the location of petroleum refineries and the organic chemicals industry. We are less sure of the relationship between SIC code 28 and the coal products industry.

A more general problem is the fact that the 2-digit manufacturing codes mix together industries with different hazard potential. For example, in the case of fabricated metals, smelting and later stage refining are combined. In chemicals, inorganic and organic industries are mixed and pharmaceuticals and primary feedstock facilities and many others are combined. Unfortunately, we had no other choice because a more refined and consistent data set extending back to at least 1950 was not available. Indeed, due to the general grouping of the data, while we felt certain that SIC codes 28 (chemicals), 34 (fabricated metals) and 35 (machinery) have high carcinogenic potential, we are less sure about the other codes we selected -- 22 (textiles), 33 (primary metal), and 37 (transportation equipment). Workers in these industries handle carcinogenic substances, but we are not sure that enough workers in those industries are exposed to classify the entire industry as high risk.

Electrical machinery, printing, leather products, and industrial and scientific instruments are industries that expose their workers to carcinogenic substances. However, the evidence was not strong enough to warrant their inclusion at this time. NIOSH's recent (1977) national occupational hazards survey suggests that these industries should have been included, and they are being incorporated into the second phase of the study.

While the census data is more general than we would have liked, it has important advantages. The most important of these is the availability of a variety of indicators such as employees, production employees, establishments, value added and others at the county scale by place of residence and by place of occupation for many years of record. The variety of indicators enabled us to make sure that our results were not based on a single indicator that may be biased or in error in one or more counties. We chose number of reporting establishments, total employees, and production employees.

Both place of residence and place of occupation data are important in a region in which intraregional commuting is common. Commuting is some counties, especially large, rural counties, may be negligible. Occupation by place of employment and residence should be nearly identical for counties that have little commuting. However, intercounty commuting is important is small, urban counties. An individual may be exposed to a substance while working in one county and residing in another. His death is recorded by place of residence. The availability of both place of residence and place of work data helped us recognize major problems caused by commuting.

Since cancer has a lengthy latency period, data beginning in 1920 or 1930 were desirable. However, due to changes in census definitions, we confined the occupational data to 1947-1972. Employment by place of work was gathered from two sources: the Census of Manufactures and County Business Patterns. The second source was used to supplement the first when the Census of Manufactures lacked the necessary information. Employment by place of residence was gathered from the Census of Population.

Problems were encountered particularly with data by workplace. Employment by workplace for 1947, 1958, 1967 and 1972 were available from the Census of Manufactures. However, the years 1948, 1959, 1967 and 1971 had to be used when County Business Patterns was used.

Other problems relating to workplace data revolve around providing data when the information was missing. The Census of Manufactures, which was the primary data source, did omit many of the less industrialized counties from its detailed county breakdown of SIC codes and associated information. As a result, the information for the omitted counties had to be obtained from the County Business Patterns. However, only the number of establishments and total employment for each SIC code could be obtained from this source. Production employment was not provided as it was for the counties analyzed in the Census of Manfactures. Production employment is important to know separately because non-production manufacturing employment includes sales, secretarial, administrative and other people who should receive less exposure to carcinogenic substances than production workers.

A method was devised for approximating the number of production employees for each county SIC code obtained from County Business Patterns. Each SIC code for each year was divided into "problem counties" and "non-problem counties". The problem counties were those in which the data source used was County Business Patterns. The non-problem counties were those in which the source used was Census of Manufactures. The second group contained both total employment figures and production employment figures. A ratio of production employment to total employment was obtained for each non-problem county in the year. After all of the ratios were calculated, a mean ratio and standard deviation were determined. The standard deviations were relatively low. Therefore, the mean ratios from the non-problem counties were multiplied by the total employment figures for the problem counties in order to estimate the production employment of the problem counties.

The mean ratios of production employment to total employment used in each calculation and resultant standard deviations are as follows:

Year	SIC	Mean Ratio	Standard Deviation
1947	22	.90	0.37

Year	SIC	Mean Ratio	Standard Deviation
1947	28	.71	.072
	33	.85	.072
	34	.84	.035
	35	.80	.029
	37	.81	.071
1958	22	.86	.047
	28	.65	.074
	33	.78	.043
	34	.76	.048
	35	.69	.097
	37	.73	.106
1967	22	.84	.045
	28	.64	.088
	33	.75	.055
	34	.79	.048
	35	.71	.068
	37	.74	.134
1972	22	.83	.064
	28	.62	.077
	33	.77	.046
	34	.76	.047
	35	.66	.069
	37	.74	.097

This procedure was not needed for the dummy SIC 99 code, all manufacturing, since all values for this code were available in the census.

Another problem in using County Business Patterns was the discrepancy between the information provided in this source and the Census of Manufactures. Often, the number of establishments or total employment for SIC codes in those counties given in both sources would vary from one source to the other. There was no regular pattern which existed; neither source was consistently high or low. It was assumed then that the information taken from County Business Patterns could cause error in the data. Nothing could be done, however, to reduce this error. The cause of the data conflict may have been the difference in data years used as the base (1947 Census and 1948 County Business Patterns).

The final problem related to using County Business Patterns was that actual values were not always given for some of the SIC codes. This was often the result of the "disclosure rule" that enabled companies to have specific information withheld from public disclosure. The information for those SIC codes in each county having a small number of establishments may be deleted. This problem had to be handled by referring to the category list in which the number of industries in a particular SIC code were broken down into employee categories. The various categories used by the County Business Patterns are as follows:

0 - 3 employees
4 - 7 employees
8 - 19 employees
20 - 49 employees
50 - 99 employees
100 - 499 employees
500 or more employees

Determining the total number of employees for a particular SIC code entailed multiplying the number of industries in each category by the median values of each category group. In the case of the 500 or more employees group, the base value was used. Unfortunately, this method used gross approximations and provided opportunity for error.

There were a few counties when even this general type of information was not available. In such cases, the Census of Manufactures was again utilized. The Census of Manufactures contains a similar type of category breakdown, but these categories are even broader and tend to be even more inaccurate. The categories used were as follows:

1 - 19 employees
20 - 99 employees
100 or more employees

The later years contained a slightly more detailed breakdown:

1 - 19 employees
20 - 99 employees
100 - 249 employees
250 or more employees

The determination of the total number of employees for each SIC code was done in a similar manner as that used with the County Business Patterns.

The final problem encountered with the occupation by place of work data was that numbers for the construction industry could only be obtained from County Business Patterns. Production employment was therefore entered as equal to total employment.

Recapitulating, establishments, total employees, and production employees data were gathered for the years 1947, 1958, 1967, and 1972. These more than 10,000 pieces of data give us a picture of the spatial distribution of jobs in the study area.

The Census of Population eliminated most of the missing data problems, but not without cost. Only total employment data were available, and they were available, only for 1950, 1960 and 1970.

The initial data were used in their raw form and converted into indices of economic specialization and geographical concentration. In general, we expected to find a positive correlation between the eleven manufacturing and construction categories and cancer and a negative association between agriculture and cancer. More specifically, the associations suggested in Table 3 were expected. The hypothesized associations are based on the literature reviewed and an analysis of the number of workers in the study area. We expected strong positive associations between the chemical industry and cancer among the male population for six body sites: liver, lung, bladder, skin, pancreas, and leukemia. Another strong association was expected between the metal industries (SIC 33, 34, 35 and 37) and lung cancer. Overall, we expected lung cancer would demonstrate the strongest and most frequent associations, followed by liver, bladder, and skin. Leukemia and pancreas would be expected to associate strongly with the presence of the chemical industry. However, given doubts about the cancer rates for these two body sites, we were not sure about the strength of the association.

Cancer of the stomach, large intestine, rectum, and kidney are sometimes mentioned in the literature. Therefore, a low positive association was thought to be possible. Finally, we did not expect any association between the industries and female breast and prostate cancer.

The multiple indicators and four time periods provided by the Census of Manufactures and County Business Patterns are important as a means of substantiating the results. We would expect the production employment data to demonstrate higher associations with cancer than the total employment figures which include office workers and the number of establishments which equally weigh small and large establishments.

Finally, we would expect the highest association between cancer and the earliest time period, 1947. The lowest associations should be with the 1972 data.

As previously indicated, the raw occupational data were transformed into more than 700 variables. The initial set of variables were total employment, production employment and establishments.

The first set of derived variables are an attempt to measure economic specialization. The assumption is that the greater the economic specialization of a region, the greater the possibility of finding the expected occupationally induced cancers. Three indicators were derived. The simplest is derived by dividing the number of people working in a specific occupation by county of residence, by the total number of residents who are employed. For example, 10,000 people who live in county X work in the metals industry in counties X, Y, and Z. One hundred thousand residents of County X work. The percentage of residents of county X who work in the metals industry in counties X, Y, and Z is 10.

TABLE 3

Expected Association Between Occupation Variables and Cancer

Operational Variable	Body Site	Result Expected
SIC 15: Construction	wm* lung, wm skin	Positive
	stomach, large intestine, rectum, liver, pancreas, wf** lung, kidney, bladder, skin	Possible low positive
SIC 22: Textile	liver, lung, bladder, skin	Positive
SIC 28: Chemical	wm liver, wm lung, wm bladder, wm skin, wm pancreas? wm leukemia?	Strong positive
	stomach, large intestine, rectum, kidney	Positive
SIC 33, 34, 35, 37: Metals	wm lung	Strong positive
	wm liver, wm skin, wm bladder	Positive
	stomach, large intestine, rectum, pancreas, kidney	Possible low positive
AGR: Farm workers	skin, bladder, liver	Possible positive
SIC 28, 34, 35: 3 Most important	wm lung, wm liver, wm bladder wm skin	Strong positive
	wm pancreas, wm leukemia	Positive
	Stomach, large intestine, rectum, kidney	Possible low positive
SIC 15, 22, 28, 33, 34, 35 37: 7 most important	wm lung, wm liver, wm bladder Stomach, large intestine, rectum, kidney, skin	Positive Possible low positive
SIC 99: All Manufacturing	wm lung	Strong positive
	stomach, large intestine rectum, liver, kidney, bladder, leukemia	Possible low positive

---

* wm = white male

** wf = white female

The second economic specialization indicator was the location quotient, a commonly used indicator of economic specialization:

$$LC_{ijt} = \frac{\frac{E_{ijt}}{E_{jt}}}{\frac{E_{irt}}{E_{rt}}}$$

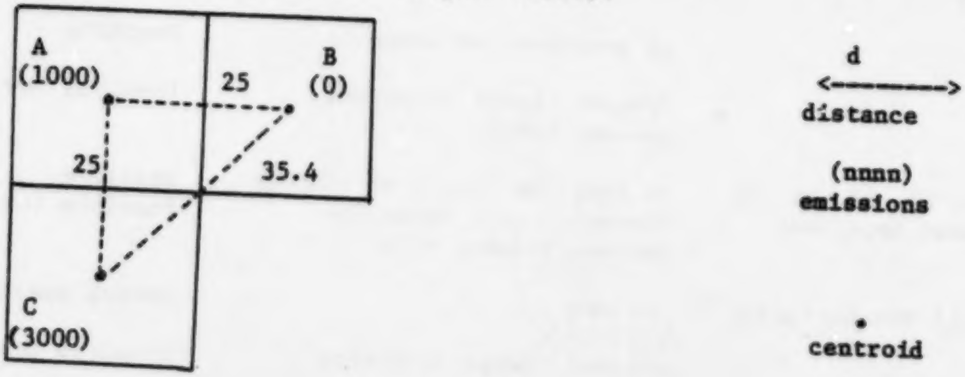
where LC_{ijt} = is the location quotient for industry i (e.g., metals) in county j (e.g., county X) in time period t (e.g., 1960).

E_{ijt} is the employment in industry i in county j in the period t;
E_{jt} is the total employment in county j in the time period t;
E_{irt} is the employment in industry i in the entire region (r) in the time period t;
E_{rt} is the total employment in region r in time period t.

The location quotient indicator assumes that the higher the location quotient, the more the economic specialization, and therefore the more the county will manifest a cancer pattern related to occupation.

The final indicator of economic specialization is a gravity potential model. The index tries to quantify the relative presence of industry in each county. A county may have little industry; if, however, it is amidst industrialized counties, industry potentially may have an important effect on the people of the county. The gravity potential model not only suggests the potential impact in the case of the non-industrialized county surrounded by industrialized counties but also quantifies the possible synergistic impact of heavily industrialized counties located adjacent to one another.

The model is best understood through a simple example. The study area has three counties: A. B, and C (see diagram below).



There are no manufacturing facilities in County B. County A has 1000 manufacturing jobs and County C 3000. The center of County A is located 25 miles from the center of County B and 25 miles from the County C. County C is located 35.4 miles from County B. If we looked only at the county totals, County C would seem to be a center of manufacturing activity and County B not a manufacturing center. However, County C's role as a manufacturing employee center is less than its 3000 employees would suggest because some of the workers commute from Counties A and B to County C and air and water emissions may leave County C. County B is affected by industry because it may have commuters to Counties A and C and may suffer some degradation of the local environment due to emissions from Counties A and C. In addition some of County B's workers may have moved from Counties C and A and now may work in retailing instead of industry. Next, some of County B's residents may have worked in industry in Counties A or C and may upon retirement have moved to County B.

In the absence of data on these important intercounty movements, this model assumes that the impact of industry on a county is directly proportional to the number of employees in that industry in all the counties and inversely proportional to the distance between them. Using the data provided above, for County A,

$$\text{Index A} = \frac{E_A}{1/2D_{AA}} + \frac{E_B}{D_{BA}} + \frac{E_C}{D_{AC}}$$

where E_A is the industrial employment in County A (1000); E_B is the industrial employment in County B (0); and E_C is the industrial employment in County C (3000); $1/2D_{AA}$ is one half the longest linear distance within County A and represents a conservative estimate of the effect of County A's industry on itself (17.7 miles); D_{BA} is the distance between the center of Counties A and B (25 miles); and D_{AC} is the distance between the centers of Counties A and C (25 miles).

$$\text{Index A} = \frac{1000}{17.7} + \frac{0}{25} + \frac{3000}{25},$$

$$= 56.5 + 0 + 120 = 176.5$$

For County B the index is:

$$\text{Index B} = \frac{E_A}{D_{AB}} + \frac{E_B}{1/2D_{BB}} + \frac{E_C}{D_{BC}},$$

$$= \frac{1000}{25} + \frac{0}{17.7} + \frac{3000}{35.4},$$

$$= 40 + 0 + 84.7 = 124.7$$

For County C the index is:

$$\text{Index C} = \frac{E_A}{D_{AC}} + \frac{E_B}{D_{BC}} + \frac{E_C}{1/2D_{CC}},$$

$$= \frac{1000}{25} + \frac{0}{35.4} + \frac{3000}{17.7},$$

$$= 40 + 0 + 169.5 = 209.5$$

The index results for this simple example suggest that the role of County B may be far more important than might be gathered from its job profile. The dimensionless gravity-potential numbers for A,B, and C are 176.5, 124.7, and 209.5, respectively.

The gravity model is a common device used to simulate the relative economic potential of regions and of commuting and migration patterns. Our version is conservative. Most models use exponents of greater than one in the denominator. In the absence of data specific to the region, we followed a guideline of avoiding exponential multiplication.

Summarizing, the gravity potential model adds an economic concentration dimension which none of the other indices can provide because they ignore the regional distribution of industry, and it is an attempt to construct a substitute for missing intraregional migration information. The gravity model index has important empirical drawbacks. The empirical drawbacks are an absence of more precise time-distance data and long-term data on intraregional movement of products and byproducts and of workers in different industries. From the theoretical perspective, the gravity model is a useful indicator of the regional economic concentration of industry viewed from the perspective of every county. However, the pollutants that accompany jobs suggest the regional economic potential may not only be a surrogate for cancer induced in the workplace but also in the ambient environment. Therefore, while we expect the gravity potential model to demonstrate higher correlations than the other occupational variables, it is a speculative model.

The next variable is employment concentration. The untransformed employment totals were divided by county area to yield an employment area index. The hypothesis underlying the transformation is that production concentrated in a small county would enhance the chance of people in the county coming into contact with any emitted carcinogenic agents.

Finally, several other variable sets were created but not used because we were not sure about their interpretation. Two of these are noteworthy. One of these variables is derived by dividing an indicator of economic activity (e.g., employment, production employment, establishments) by resident population. Counties that have a high rate would be expected to have higher rates of cancer than counties which have few jobs relative to their population. The problem with the population variable is that the expected outcome would probably not result in our region because of major differences in population size among the counties. Specifically, the literature indicates that atmospheric and water concentrations of carcinogenic substances are low. While substances might react synergistically to increase their potency and dose - response limits are not precisely known, the probability of inducing cancer in a resident at this time is thought to be low. Accordingly, whereas high populated counties like New York City and Philadelphia would be expected to have enough people to demonstrate industrially related cancers outside of the workplace, counties with populations of less than 200,000 or 300,000 might exhibit so few of these cases that they could not be distinguished from cases caused by occupational exposures and other factors.

The low concentrations of carcinogenic substances thus far measured in the ambient environment led us to design but not use a second variable set. In the next section of the chapter, we present an extremely simple air emissions dispersion model that is formed by multiplying particulate emissions by a wind rose vector and dividing the result by a measure of distance. The model was not applied to the manufacturing codes because we did not have information on the role played by each industry group in emissions. In the absence of such information, we concluded that industry's role would be more suitably reflected in the air emissions variables in which it plays an important role and in the jobs/area index which should provide an idea of local (county) effects.

Summarizing, more than 700 occupational variables were created. They range from employment totals to indicies of economic concentration and potential. Chapter 3 will describe how they were collapsed into a number more manageable for modeling.

Local Ambient Environment

Air, and to a lesser extent water and soil pollution have recently become recognized as potentially significant sources of carcinogenic and cocarcinogenic agents. The major sources of air emissions are fossil fuel generating stations, industry, incinerators, automobiles, and home heating. The major sources of water emissions are industry, generating stations, sewage treatment plants, mining operations, and runoff from urban areas and farms.

Air Pollution

The ideal data set would have air quality readings for important carcinogenic substances dating at least two decades. The historical record would be for at least two places in every county of the study area. The air quality reading would be joined by emission inventories for the same time period.

The available data bear little resemblance to the desired data. The air quality readings were quite deficient in time, place, and parameter dimensions. The emissions inventory likewise was severely restricted in time, parameters, and reliability.

After analyzing all the available data, we had to fall back to particulates as the sole indicator of air quality and emissions. Particulates is an all encompassing term for a multitude of different forms of particulate matter from natural and man-made sources. The air quality term used when speaking of paraticulates is total suspended particulates (TSP). The national ambient air quality standard for TSP is 75 ug/m^3 (annual geometric mean) and 260 ug/m^3 for a twenty-four hour period.

Labelle (1955) was one of the first investigators to report that particulates were carriers of toxic materials. Cloudiness around cities may be induced by particulate production. Measurements have consistently confirmed that particulates, which act as condensation nuclei, are more numerous by one or two orders of magnitude in urbanized regions that in the countryside (Landsbeg, 1970). Particulates can be associated with the formation of harmful aerosols. De Maio (1966) collected respirable and nonrespirable particulates in the air of Pittsburgh, Pennsylvania. He found that more than 75 percent of the polycyclic aromatic hydrocarbons on particulates were associated with respirable particles. A number of toxic metals and chemicals, including benzo(a)pyrene, predominate in the small particles in most urban aerosols (Natusch, 1974). This occurs because small particles act as condensation nuclei for irritant gases or carcinogenic hydrocarbons. These small particles are capable of being deposited in the lung where they remain for weeks or even years (Natusch, 1974). Deposition of these particles usually occurs at the bronchial bifurcation and in the lower airways. Relative concentration depends on particle size, density, shape, and tendency to aggregate and impact upon the lining of the lungs and upper respiratory tract (Walbott, 1973). Lee (1972) for example, has shown that particulates in automobile exhaust are extremely small, consequently most of them can be inhaled and retained. Overall, while particulates may be a good indicator of the presence of some carcinogens in the air, TSP is certainly not the only or even the most desirable indicator.

Ambient air concentration data were sought for the earliest time period possible, subject to the following constraints:

1. comparability to data collected for later time periods; and
2. adequate number of data points throughout the study area during a given time period.

Because the time between exposure to carcinogenic agents and mortality due to cancer is generally long, we sought the earliest available air pollution data for the forty-nine county study area and environs. The available air pollution data are not extensive. While indicators of particulate concentrations in New York City had been gathered as early as the 1940's, similar data were not available for the other counties. The first set that approaches regional coverage is available from the U. S. Department of Health, Education and Welfare starting in 1953 (U.S. Public Health Service, 1958).

The first air pollution data set used in this study is for the period 1953-1957. The air quality parameter used is total suspended particulates measured in micrograms per cubic meter. An arithmetic average of the five year period was the only variable we could calculate. Since the later data are geometric means, the earlier and later data sets are not directly comparable. When the National Air Sampling network was initiated in 1953, the number of air monitoring stations was severely limited. Data are available for 31 air monitoring stations within our 49 county study area for the period 1953-1957 (Table 4). In addition, four sites outside the region were used because they were closer to some of the counties than any of the stations in the study area. Most of these stations operated for less than the full five year period and were located in the urban centers of our study area. Philadelphia and New York City accounted for eleven of the thirty-five air monitoring stations used in this data set. The following stations outside the study area were added to the 1953-1957 data set: Hartford, Connecticut; Pittsfield, Massachusetts; Schenectady, New York; and Scranton, Pennsylvania.

To determine if the spatial distribution of cancer mortality is related to air pollution, we had to know air quality in each of the counties of our study area during the period 1953-1957. This was a problem due to the small number of air monitoring stations and their concentrated distribution. A mathematical method of estimating the exposure of the population of each county to air pollution was used to simulate missing values.

A method developed by U.S.E.P.A. (1976) to estimate the exposure of populations to air pollution recommends a pseudo-parabolic function to derive a pollutant concentration at a specified point.

A receptor point is chosen at which the exposure to air pollution will be determined. The relative distance from the receptor point to the three nearest air monitoring stations and the concentrations of pollutant at those three air monitoring stations are used to calculate the concentration at the receptor point.

$$CRP = \sum_{i=1}^{3} (C_i/r_i^2) \Big/ \sum_{i=1}^{3} (1/r_i^2) \qquad \text{for } r_i \neq 0$$

where CRP = estimated concentration at the receptor point
C = concentration at the air moitoring station
r = distance from the air monitoring station to the receptor point

TABLE 4

Location Of Air Monitoring Stations For Which We Collected Data
1953-1957, 1972-1975

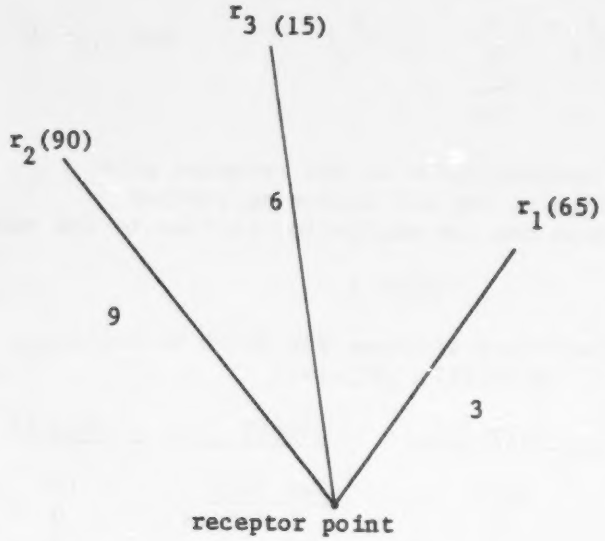
COUNTY	1953-57	1972-75
Connecticut	(4)	(45)
Fairfield	1	19
Litchfield	1	6
New Haven	2	20
Delaware		
New Castle	(1)	(11)
New Jersey	(12)	(72)
Atlantic	1	2[a]
Bergen	0	5[a]
Burlington	0	3
Camden	1	3
Cape May	0	2
Cumberland	0	0
Essex	2	6
Gloucester	1	3
Hudson	1	5[a]
Hunterdon	1	1
Mercer	1	4[a]
Middlesex	1	12
Monmouth	1	5[a]
Morris	0	3[a]
Ocean	1	5
Passaic	0	2[a]
Salem	0	1
Somerset	0	3
Sussex	0	1
Union	1	4
Warren	0	2[a]

COUNTY	1953-57	1972-75
New York	(4)	(45)
Dutchess	0	3[a]
Nassau	0	8[a]
New York City	4	5[a]
Orange	0	1
Putnam	0	2
Rockland	0	3
Suffolk	0	5[a]
Sullivan	0	0
Ulster	0	6[a]
Westchester	0	12[a]
Pennsylvania	(10)	(23)
Bucks	1	5
Carbon	0	0
Chester	0	4
Delaware	1	2
Lehigh	1	2
Monroe	0	0
Montgomery	0	5
Northampton	0	4
Philadelphia	7	1[a]
Pike	0	0

---

[a] Multiple sites at one or more locations.

The following example illustrates the method. Air monitoring stations r_1, r_2 and r_3 have concentrations of 65, 90, and 15 ug/m^3 of total suspended particulates, respectively.



$$CRP = (65/3^2 + 15/6^2 + 90/9^2)/(1/3^2 + 1/6^2 + 1/9^2),$$

$$= (7.2222 + .4167 + 1.1111) / (.1111 + .0278 + .0123) = 8.75000/.1512 = 57.91$$

In this example there is no air monitoring station sufficiently proximate to the receptor point so that its recorded level of air pollution can be assumed to be a good estimate of the concentration of pollutants at the receptor point. The method described computes a concentration of 58 ug/m^3 at the receptor point.

Each of the forty-nine counties in our study area was assigned from one to three receptor points depending on the population of that county. It was assumed for the purposes of the study that the population of the entire county was exposed to the concentration of pollutant calculated for those receptor points. This assumption was necessitated by the very limited air pollution data available. A single arithmetic mean value for air pollution for each was computed.

The results of applying the EPA exposure model required some modification in our study area. The EPA method produces reasonable results only when at least one air monitoring station was in close proximity to the population receptor point of interest. How close the station must be to produce reasonable results will vary from site to site. In our study area, based on empirical tests, twenty miles was found to be the upper limit. The formula uses distance so that the concentrations of pollutants at far removed air monitoring stations have only marginal impact on the concentration of a local station.

Since the EPA method calculates a concentration based on the relative distances of the three nearest stations, it does not recognize the decrease in concentration of total suspended particulates that normally occurs over absolute distance.

For example the EPA method could calculate a concentration of total suspended particulates for a rural county in northeastern Pennsylvania based on concentrations found in air monitoring stations in the urban core of New Jersey. The relative distance of the three stations determines the concentration, but the results reflect the air quality determined at the air monitoring stations.

The lack of rural air monitoring stations causes severe problems when trying to study the effects of air pollution on rural populations. Any method that attempts to calculate the concentration of pollutants in a rural area based on concentrations in an urban area will encounter difficulties. The EPA method used in our study suffered from this problem. Only two of the thirty-five air monitoring stations in the period 1953-1957 were characterized as other than urban.

Several steps were taken to modify the unreasonable concentrations obtained for certain rural counties by the EPA method. The basis of this modification was a relaxation of the EPA method rule that the three closest air monitoring stations be used to calculate the population exposure to air pollution. One or more air monitoring stations were substituted for the closest three stations when those stations produced unreasonable results determined by knowledge of the area in question and the region in general.

The substitutions were made (e.g., nearest rural station for nearest urban station) or terrain (e.g., nearest rural station in same physiographic province), or some combination of factors. The most common modification of the EPA method in our study area was the replacement of one or more of the closest three air monitoring stations by a station in a location known to resemble the population receptor points in question in relative urbanization. For many receptor point concentrations, a non-urban air monitoring station replaced an urban one in the calculation, even though it was more distant from the receptor point than the urban station.

Knowledge of the prevailing atmospheric conditions in the study area suggested another modification. Since concentrations of total suspended particulates fall off with distance, wind patterns in the study area have a profound influence on the dispersion of pollutants. Since the prevailing winds in our study area are from the southwest and northwest, some air monitoring stations in this quadrant replaced stations located northeast of receptor points when the concentration of total suspended particulates were calculated.

The modifications to the EPA method improved some of the glaring problems with the 1953-1957 air quality data set. After the modifications, several rural counties had lower concentrations of total suspended particulates that had originally been calculated. While the modifications caused an improvement, the data set remains weak. The limited air pollution data for the period 1953-1957 cannot be overcome by technique. The data set can only be as good as the data collected.

Given the impetus from the various Clean Air Acts legislated since 1963, during the mid-1960's, the states began to develop their own system for monitoring pollutant concentrations in the ambient air. New York State was the first to establish a statewide monitoring system, doing so in 1964. The total number of monitoring sites increased into the 1970's. New Jersey began setting up their existing monitoring system in 1966. The City of New York began an extensive monitoring program (35 sites) in 1969. The State of Connecticut began their program in 1971, the Commonwealth of Pennsylvania and the State of Delaware in 1972. The City of Philadelphia measured ambient concentrations of particulates at two sites until 1974 when the network was increased to 12 sites.

The time period of 1972-1975 was chosen as the second data set because of the lack of data for Pennsylvania and Delaware before that time and because of our desire to average the available data over some consistent time period.

The geographical distribution of particulate monitoring sites (that now use high volume sampling devices almost exclusively) is depicted in Table 4. It illustrates that while the number of sites has increased since the 1953-1957 time period, they generally remained concentrated in the urban-industrial areas of the region. A number of the rural counties in the study area still have no monitoring sites.

The data available for the 1972-1975 time period took the form of annual geometric means for particulate concentration (in ug/m^3). Four year averages of these geometric means were taken as a general indicator of particulate concentration at each site. In instances where data were not gathered for each of the four years, averages were taken for the years in which measurement did take place, with the exception of those sites at which measurements were taken during only one year of this period. In this case the data from those monitoring sites were not utilized in this process.

While less a problem than the 1953-1957 data set, as noted above, a number of counties, such as Pike, Carbon, Monroe, and Sullivan had no monitoring sites within their boundaries. As a result, we had to estimate the air quality in these counties on the basis of measurements taken in other counties in the region that have similar rural characteristics.

The technique used to obtain an air quality indicator for each county over this time period was a relatively simple one. In the urban counties, where there were generally at least three monitoring sites that provide usable four year averages of geometric means, the average of all sites in the county was used as the air quality indicator.

In the rural counties that had only one or no monitoring sites, data from nearby sites in other rural counties were substituted. This was our only alternative, given the composition of the data base.

In addition to being limited to a single parameter and having to simulate values for many counties, another serious problem has to be noted. A number of monitoring locations were changed within the boundaries of a minor civil division, thus presenting problems of comparability. For example, one site registered an annual geometric mean of particulate concentration of (142)ug/m^3 in 1973. That site was phased out after 1973 and replaced by a site within the confines of the same town that registered an annual geometric mean of (54)ug/m^3 in 1975. Accepting these data on face value would result in the conclusion that the region improved dramatically during one year. Such improvement is clearly not the case. The major portion, if not all, of the difference can be attributed to the change in the location of the site.

It is beyond the scope of the study to attempt to determine the reasons for these changes of location, although they are certainly not irrelevant. The reasons may differ from site to site. Notwithstanding these reasons, however, such changes, if they continue in the future, are likely to impair the accuracy of longitudinal studies that utilize the data from such sites.

The resultant air quality indicators are presented on Table 5 below. In viewing the table for 1972-1975 one can see that only New York City, Philadelphia, and Hudson County are at or exceed the Primary Ambient Air Quality Standard of 75 ug/m^3, and only 12 exceed the secondary standard of 60 ug/m^3. This is to be expected, however, since the indicator is a result of averaging all sites in the county. Many counties that are generally urbanized have large areas that are predominantly low density residential in nature and measurements in these areas will provide a lower limit to the range of particulate concentration measurements in a county.

In conclusion, we have a single indicator, TSP, for every county in the study area for the periods 1953-57 and 1972-75. We expected correlations between TSP concentration and lung cancer and lesser correlations with cancers of the digestive tract. However, the data are so dubious for the rural counties that we were not confident of a strong relationship. Within the study area, the New Jersey counties had the densest set of data in and surrounding the state. Accordingly, relationships within New Jersey should be the most reliable of those of any state included in this study. Overall, the air quality data are poor in comparison to the socio-economic data.

TABLE 5

TSP Air Quality Indicator
Average Concentration (ug/m^3)

County	Arithmetic Mean 1953-57	Geometric Mean 1972-75
Fairfield, Conn.	112	57
Litchfield, Conn.	73	42
New Haven, Conn.	120	58
New Castle, Del.	189	63
Atlantic, N.J.	114	46
Bergen, N.J.	168	65
Burlington, N.J.	135	39
Camden, N.J.	191	40
Cape May, N.J.	116	28
Cumberland, N.J.	135	32
Essex, N.J.	133	65
Gloucester, N.J.	137	36
Hudson, N.J.	182	81
Hunterdon, N.J.	59	41
Mercer, N.J.	148	50
Middlesex, N.J.	96	56
Monmouth, N.J.	78	42
Morris, N.J.	112	40
Ocean, N.J.	62	34
Passaic, N.J.	145	57
Salem, N.J.	174	67
Somerset, N.J.	86	47
Sussex, N.J.	63	35
Union, N.J.	103	60
Warren, N.J.	63	44
Dutchess, N.Y.	78	45
Nassau, N.Y.	171	54
New York City, N.Y.	968	78
Orange, N.Y.	87	47
Putnam, N.Y.	110	43
Rockland, N.Y.	161	50
Suffolk, N.Y.	151	46
Sullivan, N.Y.	75	34
Ulster, N.Y.	75	61
Westchester, N.Y.	156	53
Bucks, Pa.	126	52
Carbon, Pa.	107	39
Chester, Pa.	156	68
Delaware, Pa.	135	73
Lehigh, Pa.	111	54
Monroe, Pa.	69	39
Montgomery, Pa.	196	70
Northampton, Pa.	101	64
Philadelphia, Pa.	227	75
Pike, Pa.	115	35

A second data set utilized in studying the relationship between air pollution and cancer mortality was the emission inventories by county that were developed by the states as required by the Clean Air Act of 1970.

Emission inventories for particulates were obtained for the study area from the appropriate state agencies. Two major problems were found: the Pennsylvania data and the dates of the inventories. The inventories were analyzed for any glaring discrepancies in the data developed by each state. Pennsylvania's emission inventory was found to bear little relationship with those from the other states in the study area and to other existing sources. Some counties had been assigned emission values that seemed impossible given their economic and demographic structures. Representatives of Pennsylvania recognized the inadequacy of the data set. They have been preparing a new inventory. However, it was not completed and could not be utilized by the study in its present form. The City of Philadelphia supplies their own inventory. The remaining nine Pennsylvania counties were the problem. The weakness of the Pennsylvania data was a serious blow. We had hoped to use the emissions inventory disaggregated into industrial, energy and other categories. The absence of useful Pennsylvania data eliminated that goal and made us fall back to total emissions as a variable.

Another potential problem with the overall data set was that the emission inventories for New York, New Jersey, Pennsylvania, and Delaware represented 1970 emissions. The data from Connecticut were estimates for 1973.

It was decided to attempt to salvage the Pennsylvania and Connecticut data through the use of linear regression analysis. We hypothesized that the emissions data submitted by all states except Pennsylvania could be plotted against population and employment in the non-Pennsylvania counties with a resultant linear relationship. Plotting the data resulted in what appeared to be a linear relationship between particulate emissions and both employment and population. A correlation analysis resulted in a value very near .90 in both cases. This result, however, was thought to be biased by the overriding determination of the equation by the City of New York. Removing the New York City data from the analysis resulted in a change in the correlation, thus confirming our suspicions.

We then decided to perform a log/log transformation on the data and rerun the regression analysis with the resultant transformed data. The log-arithemic transformation reduces the impact of New York City and Philadelphia. Removing the 1973 Connecticut data from the analysis did not significantly alter the results, so they were kept in on the final run.

The r^2 value between log of emissions and log of population is .812, between log of emissions and employment is .808. The equation was highly significant and unbiased. Accordingly, the log of emissions with log of population equation was used to predict emissions for the nine Pennsylvania counties (excluding Philadelphia) with unrealistic emissions inventories. This gave us a complete data set on particulate emissions for the 45 places. This is presented in Table 6.

TABLE 6

Particulate Emissions 1970

County	Emissions, Tons/Year
Fairfield, Conn.	8831
Litchfield, Conn.	1489
New Haven, Conn.	12722
New Castle, Del.	39140
Atlantic, N.J.	2414
Bergen, N.J.	16203
Burlington, N.J.	5197
Camden, N.J.	6013
Cape May, N.J.	2511
Cumberland, N.J.	6492
Essex, N.J.	14449
Gloucester, N.J.	4043
Hudson, N.J.	11110
Hunterdon, N.J.	1234
Mercer, N.J.	9133
Middlesex, N.J.	23258
Monmouth, N.J.	6681
Morris, N.J.	4831
Ocean, N.J.	2884
Passaic, N.J.	5163
Salem, N.J.	2602
Somerset, N.J.	6458
Sussex, N.J.	742
Union, N.J.	9314
Warren, N.J.	1409
Dutchess, N.Y.	5642
Nassau, N.Y.	26516
New York City, N.Y.	114066
Orange, N.Y.	5011
Putnam, N.Y.	575
Rockland, N.Y.	3764
Suffolk, N.Y.	19579
Sullivan, N.Y.	558
Ulster, N.Y.	3897
Westchester, N.Y.	13467
Bucks, Pa.	7946
Carbon, Pa.	1167*
Chester, Pa.	5436*
Delaware, Pa.	11164*
Lehigh, Pa.	5525*
Monroe, Pa.	1077*
Montgomery, Pa.	11915*
Northampton, Pa.	4579*
Philadelphia, Pa.	35098*
Pike, Pa.	278*

*Estimated via regression equation

Sources: Emission Inventories from the State Implementation Plans for New York, New Jersey, Connecticut, Delaware and other data from the City of New York and the City of Philadelphia

This particulates emissions vector was divided by the county land area vector to produce an emissions density variable. We expected the emissions density variable would demonstrate stronger relationships than the untransformed emissions variable.

The third air emissions variable is an attempt to simulate the potential impact of emissions in one county on other counties. Long distance movement of emitted particles and gases should be expected in the study area. A concentrated of industrial, residential, commercial, transportation, energy and other emitters runs through the center of the region. The dominant winds come from the southwest and northwest and can carry pollutants from both the Great Lakes and east coast megalopoli through the region. While most of the region is in the flat coastal plain, the foothills of the Appalachians lie to the west of the region and can trap contaminants in the region.

Empirical evidence supports the theoretical arguments that pollutants move along the central spine of the region. Cleveland and associates (1976) describe the movement and formation of ozone in the region from Philadelphia to northern Massachusetts. Ozone readings were taken at 32 sites and meteorological readings of eight sites from May to September 1974. Ozone was traced from New Jersey to northeastern Massachusetts. The highest readings were found in southwestern Fairfield County, Connecticut.

Given the crudeness of the emissions data, we certainly could not justify the use of a sophisticated diffusion model; however, we were not willing to forego any attempt to simulate the movement of particulates. Therefore, we devised a simple model which recognizes two of the three major factors that should influence the diffusion of effluents: distance and wind direction. The third factor is the presence of physical barriers that would impede the free movement of air masses. The study area has few such barriers, especially in the urban east.

The distance and meteorological components were included by assuming that the impact of emissions is inversely proportional to the distance between the counties and directly proportional to the proportion of the time the wind blows from one county to another. Straight-line distance between the centers of the counties was used as the distance component.

The wind direction component required more complex calculations. Meteorological records were supplied by the United States Weather Service. The records summarize meteorological data. Dr. Donald Dunlap, who served as a meteorologist for the State of New Jersey and teaches meteorology, indicated which weather service stations were representative of the 49 counties (Table 7). The years of record included in the Weather Bureau reports vary from five years for Binghamton to more than twenty years for Atlantic City, Newburgh, and Lakehurst (Table 8).

TABLE 7

Counties by Wind Station

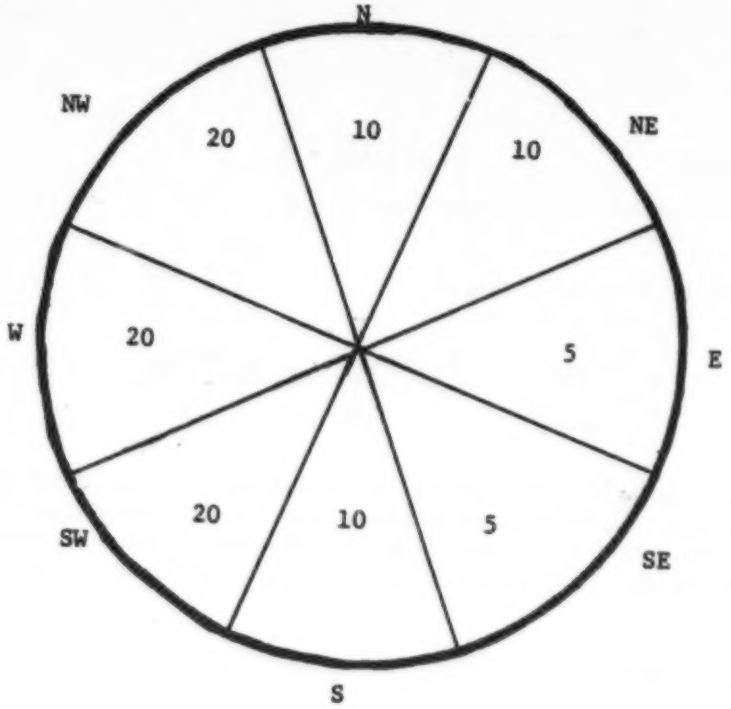
County	Weather Station
Fairfield	Hartford
Litchfield	Hartford
New Haven	Hartford
New Castle	Wilmington
Atlantic	Atlantic City
Bergen	Newark
Burlington	Philadelphia
Camden	Philadelphia
Cape May	Atlantic City
Cumberland	Wilmington
Essex	Newark
Gloucester	Wilimington
Hudson	Newark
Hunterdon	Allentown
Mercer	Trenton
Middlesex	average of Trenton and Newark
Monmouth	Lakehurst
Morris	Allentown
Ocean	Lakehurst
Passaic	Allentown
Salem	Wilmington
Somerset	average of Trenton and Newark
Sussex	Allentown
Union	Newark
Warren	Allentown
Dutchess	Newburgh
Nassau	New York City
New York City	New York City
Orange	Newburgh
Putnam	Newburgh
Rockland	Newburgh
Suffolk	New York City
Sullivan	Binghamton
Ulster	Newburgh
Westchester	Newburgh
Bucks	Trenton
Carbon	Allentown
Chester	Philadelphia
Delaware	Philadelphia
Lehigh	Allentown
Monroe	Scranton
Montgomery	Philadelphia
Northampton	Allentown
Philadelphia	Philadelphia
Pike	Scranton

TABLE 8

Air Station Data

Station	Location	Years of Record
New York City	J.F.K. Airport	1951-60
Newark	Newark Airport	1951-60
Philadelphia	Philadelphia Airport	1951-60
Hartford	Hartford Airport	1951-60
Atlantic City	Midtown	1947-69
Allentown	Airport	1949-54
Newburgh	Stewart A.F.B.	1942-69
Scranton	Airport	1951-60
Wilmington	Airport	1951-60
Trenton	McGuire A.F.B.	1945-70
Lakehurst	Naval Air Station	1942-69
Binghamton	Airport	1956-60

Wind roses were constructed from the Weather Bureau data reflecting the following eight cardinal points: north, northeast, east, southeast, south, southwest, west and northwest. The average annual proportion of the time that the wind blows from one of the directions was recorded for each of the eight directions. A simplified example is illustrated below.

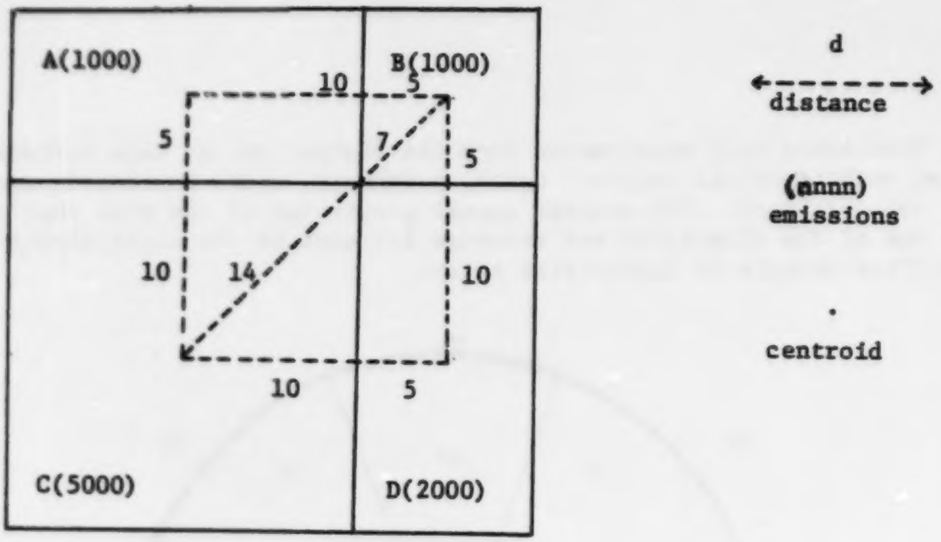


The wind blows from the southwest twenty percent of the time. It blows from the east only five percent of the time. It blows from the north ten percent of the time.

The appropriate wind rose is then placed over each county. The percentage of the time the wind blows from the direction is multiplied by the emissions/distance ratio for each county.

A simple example will illustrate the model.

A study area has four counties: A, B, C, D. Their emissions and distances between their centroids are indicated on the next page.


A(1000)
B(1000)
C(5000)
D(2000)
10
5
5
7
5
10
14
10
10
5
d
distance
(nnnn)
emissions
centroid

The calculation for B is the following:

$$\text{Index B} = \frac{E_{A}(WR\%_{AB})}{D_{AB}} + \frac{E_{A}(WR\%_{BB})}{1/2D_{BB}} + \frac{E_{A}(WR\%_{CB})}{D_{CB}} + \frac{E_{A}(WR\%_{DB})}{D_{DB}}$$

where Index B is the dimensionless index; the E values are the emissions from each county; the D values are the distances between the county centroids; WR% are the proportions of the time the wind blows from the source of the emissions to the place of impact; the factor 1/2D for the county under consideration was described earlier.

$$\text{Index B} = \frac{1000(.20)}{15} + \frac{1000(1.00)}{7} + \frac{5000(.20)}{21} + \frac{2000(.10)}{15} =$$

$$13.33 + 142.86 + 47.62 + 13.33 = 217.14$$

County A emits 1000 tons of particles at an average distance of 15 miles from the centroid of County B. When the sample wind rose is placed on the centroid of B, A is found to be west of B. The westerly wind blows 20 percent of the time.

County B emits 1000 tons at an conservative distance of 7 miles from its centroid. The emissions affect it 100 percent of the time.

County C emits 5000 tons at an average distance of 21 miles from B. C is predominantly in the southwest quadrant and receives the effect 20 percent of the time.

Finally, County D emits 2000 tons 15 miles from B. County D is south of B, and therefore its emissions only effect B 10 percent of the time.

The calculations for County C are shown below to illustrate the attributes of the model:

$$\text{Index C} = \frac{E_{A(}WR\%_{AC)}}{D_{AC}} + \frac{E_{B(}WR\%_{BC)}}{D_{BC}} + \frac{E_{D(}WR\%_{CC})}{1/2D_{CC}} + \frac{E_{A(}WR\%_{DC})}{D_{DC}}$$

$$\text{Index C} = \frac{1000(.10)}{15} + \frac{1000(.10)}{21} + \frac{5000(1.00)}{24} + \frac{2000(.05)}{15} =$$

$$6.67 + 4.76 + 357.14 + 6.67 = 375.24$$

The emissions of County C are five times (5000/1000) the emissions of County B. However, the distance-wind rose index value of C is only 1.73 times (375.24/217.14) the value of B while the west winds carry a relatively large proportion of the emissions of C and A to B, the affect of B, A, and D on C is slight.

While we believe the model may be a better abstraction of reality than the simple county totals and fits the quality and quantity of the data we had available, it has a variety of shortcomings. It assumes that particles fall to the earth in a linearly proportional rate. Most studies suggest that the fall-off is far more rapid for large particles and much less rapid for small particles. However, without precise data, we used a simple linear model to avoid the compounding of error pitfalls of exponential models. Second, the centroids of the counties are used as the points from which to measure distance and to choose a wind rose direction. Third, emissions from outside the region are not included in the model. Fourth, the annual average emissions are weighed by annual average wind directions. In fact, a substantial proportion of emissions of some particles and gases may occur during a specific season, and therefore the annual wind rose proportions may be inappropriate. Finally, and most important, the distance-wind rose index model is an abstraction of what we think may be happening to carcinogens emitted into the atmosphere, not what we know is happening. While we have used it, we regard it as a very speculative abstraction.

Given these limitations and strong cautions, we expected the distance-wind rose model to produce high correlations with lung cancer and lower correlations with digestive tract cancers. However, given the facts that the data reflect emissions after the 1950-1969 study period, are simulated for nine of the forty-five units, and are limited to particulates, we sought to collect separate data on two activities that are responsible for a large proportion of the region's emissions: fossil fuel power generating stations and motor vehicles. Generating stations emit from relatively few points. Motor vehicles are ubiquitous in the study area. We assumed that correlations with the generating station and motor vehicle data would corroborate the results of correlations with the emissions data.

Electric Generating Stations

Electric generating stations burning fossil fuels are the source of about half of the sulfur oxides, one-fourth of the particulates, one-fourth of the nitrogen oxides, and lesser proportions of the carbon monoxide and hydrocarbons, including benzo(a)pyrene, emitted into the atmosphere in the United States (Energy Policy Staff, 1968). Fossil fuel electric generating stations may not be as important a source of carcinogenic organic compounds as cigarettes, autos, and home heating. They are, however, a signifcant source. In addition, the fact that the organics are emitted from high stacks in combination with other pollutants suggests that their role may be greater than the proportion of carcinogenic pollutants they represent. The release of organics and other chemically actively pollutants from high stacks can lead to the formation of dangerous aerosol mists that increase the potential for both short and long-term harmful health effects. Finally, in regions that use oil as the major fuel (such as the study area), the harmful potential is enhanced because oil produces smaller and more irregularly shaped particulates than coal. Overall, fossil fuel emissions may be directly carcinogenic and may harm the respiratory and digestive systems, thereby rendering them more susceptible to other carcinogens.

Our goal was to measure annual emissions by fossil fuel stations and correlate these emissions with cancer mortality. Emission data are available for recent years, but not for a long enough historical period. Annual tons of fossil fuels burned for every generating stations are available from 1938 to the present. These data constituted the base of our electric generating station variables.

In 1938, there were 20 generating stations located in 13 of the 49 counties. By 1972, there were 46 generating stations located in 28 of the 49 counties. The fossil fuel fired facilities not only spread and increased in number, but also greatly expanded in size. Average plant size was 102 MW in 1938 and 230 MW in 1960. The new facilities are usually considerably in excess of 500 MW.

Annual burn of fossil fuels data were gathered from the Federal Power Commission's (FPC) S-72 publication "Steam Electric Plant Construction Cost and Annual Production Expenses: 1938 to 1947" and from annual supplements. The FPC received the information in annual reports that the utilities are required to furnish. Summations are published by the United States Government Printing Office.

Statistics for both privately and publicly owned steam-electric plants are included in the continuous record with two exceptions:

A. By-product steam-electric plants have always been excluded because "the operations and resultant costs are not generally comparable with the costs of conventional electric power production operations". These plants deliver processing steam for industrial purposes as well as electric power.

B. Plants with an installed generating capacity of 25 MW or less have been excluded since 1965.

Two plants in the study area fall into category A. They are Atlantic City Electric Company's Deepwater Plant in Salem County (supplying steam to the DuPont plant at Deepwater) and Public Service's Linden Plant in Union County (supplying steam to Exxon at Bayway). We obtained data for both of these from the annual reports of the two utilities. Only one plant, the Housatonic Avenue Plant in Derby, Connecticut, operated by Housatonic Public Service appeared to fall in category B. Following a merger of Housatonic Public Service with Connecticut Power and Light, the Housatonic Avenue Plant was phased out in 1961.

The form in which the data are collected and presented has changed little since 1938. The main difference, the disaggregation of cost data in recent years, does not affect our purpose. Another change, of more interest than importance in this case, is the addition in 1960 of the condensing water supply source for each plant.

In summary, the data base for the electric utility industry has grown increasingly comprehensive over the years. It is our most reliable ambient environmental data set.

The quantities of each of the fuels consumed are interesting if only to emphasize the gross yearly increases. Conversion of tons of coal, barrels of oil and cubic feet of gas consumed to the total BTU content of each provides a means of comparison among the three. For example, Philadelphia Electric Company's Chester plant, with a 256 MW capacity as of 1972, generated 1036.0 million KWH in 1970. To generate that amount of electricity required 43.0 tons of coal at 11,665 BTU/pound; 1,859 million barrels of oil at 148,219 BTU/gallon and 2.834 billion cubic feet of gas at 975 BTU/cubic foot. Each KWH produced required on the average, 14,806 BTU. If each of the quantities of fuel is converted to its BTU content, this can in turn be converted to a percentage of the total BTU used to produce, in this case, 1036.0 million KWH. For plants using a combination of two or all three of fossil fuels, a chronological series of these percentages highlights the pattern of use of the pertinent fossil fuels.

For our purposes a better means of comparison is to convert the BTUs produced by two of the fuels to equivalents of the third. The most common conversion, and the one applied in the FPC annual summaries to national totals of fuel burned, is the conversion of oil and gas to coal equivalents. A footnote states that this conversion is based on generation and not on BTU content of the fuels.

Any difference between this concept and the method described below by which we achieved the conversions seems nominal as the quantity of fuel consumed and the BTU content of said fuel are directly related to the amount of electricity generated. The basic information we used was given in the following form:

FUEL	QUANTITY OF FUEL CONSUMED	BTU CONTENT
Coal	1,000's of 2,000 lb. tons	BTU/lb.
Oil	1,000's of 42 gal. bbls.	BTU/gal.
Gas	1×10^6 cubic feet	BTU/c.f.

Determination of total BTUs for each type of fuel was straight forward:

(1) Coal: (1,000 tons) (2,000 lb/ton) (BTUs/lb) = BTUs
e.g.: (715.1×10^3 tons) (2×10^3 lb/ton) (12,832 BTU/lb) =
$18,352,326 \times 10^6$ BTU

(2) Oil: (1×10^3 bbls) (42 gal/bbl) (BTU/gal) = BTUs
e.g.: ($2,835.7 \times 10^3$ bbls) (42 gal/bbl) (139,827 BTU/gal) =
$16,653,311.8 \times 10^6$ BTU

(3) Gas: (1×10^6 cf) (BTU/cf) = BTUs
e.g.: ($10,550.9 \times 10^6$ cf) (1,060 BTU/cf) = $11,183,954 \times 10^6$ BTU

To convert the oil and gas BTUs to coal equivalents, or in other words make the oil and gas equivalent to tons of coal, requires a conversion factor which is the BTU content of a ton of coal in any given year:

(4) (2,000 lb/ton) (BTU/lb) = BTU/ton
e.g.: (2,000 lb/ton) (12,832 BTU/lb) = $25,664 \times 10^3$ BTU/ton

Thus the 2,835.7 bbls of oil in (2) above become the equivalent, in heat content, of:

$$(5)\quad \frac{16,653,311.8 \times 10^6 \text{ BTU}}{25,664 \times 10^3 \text{ BTU/ton}} = 648.90 \times 10^3 \text{ tons coal}$$

and the $10,550.9 \times 10^6$ cf of gas in (3) above become the equivalent, in heat content, of

$$(6)\quad \frac{11,183,954 \times 10^6 \text{ BTU}}{25,664 \times 10^3 \text{ BTU/ton}} = 435.78 \times 10^3 \text{ tons coal}$$

For those years when no coal was consumed, the BTU content of coal in the last year it was burned was used, and for those few plants which had never burned coal, a standard 25,000 BTU was used.

The 1938-1972 time period was broken into three parts, 1938-1947, 1950-1965, and 1970-1972 reflecting the role of coal as a fuel. In the early period, coal was the only fuel used in the vast majority of the generating stations. The period 1950-1965 witnessed the major introduction of oil as a fuel. With the passage of the Clean Air Act in 1970 and restrictions against the burning of high sulfur fuels, oil and gas have dominated the New Jersey and New York generating system during the 1970's.

Eight initial electric generating station fossil fuel variables were developed for every county using the above methods and data:

1. Total tons coal burned, 1938-1947
   (henceforth abbreviated TC)

2. TC, 1950-1965

3. TC, 1970-1972

4. TC, 1938-1972

5. Total tons coal equivalent burned, 1938-1947 (henceforth abbreviated TCE)

6. TCE, 1950-1965

7. TCE, 1970-1972

8. TCE, 1938-1972

Following the assumption that emissions spread over a small county are more harmful than the same emissions spread over a large county, the above eight variables were divided by the area of every county to produce eight electric generating stations, emissions concentration variables:

9. TC, 1938-1947/square mile

10. TC, 1950-1965/square mile

11. TC, 1970-1972/square mile

12. TC, 1938-1972/square mile

13. TCE, 1938-1947/square mile

14. TCE, 1950-1965/square mile

15. TCE, 1970-1972/square mile

16. TCE, 1938-1972/square mile

Next, the coal and coal equivalent variables were transformed with the wind direction/distance model.

17. (TC, 1938-1947) (wind direction vector)/(distance vector)

18. (TC, 1950-1965) (wind direction vector)/(distance vector)

19. (TC, 1970-1972) (wind direction vector)/(distance vector)

20. (TC, 1938-1972) (wind direction vector)/(distance vector)

21. (TCE, 1938-1947) (wind direction vector)/(distance vector)

22. (TCE, 1950-1965) (wind direction vector)/(distance vector)

23. (TCE, 1970-1972) (wind direction vector)/(distance vector)

24. (TCE 1938-1972) (wind direction vector)/(distance vector)

In general, we expected a positive correlation between the twenty-four variables and cancer, except for male prostate and female breast cancer. In particular, we expected a strong positive association with lung cancer and weaker positive correlations with some of the digestive tract variables.

The eight total tons of coal variables are the weakest variables. They should be expected to produce the lowest correlations because they omit oil and gas and do not take into account the diffusion of emissions. The total tons of coal equivalent transformed by the wind rose and distance variables should produce the most realistic associations. The 1938-1947 period should be expected to produce the highest correlation because it provides the longest latency period. However, there were relatively few generating plants during the 1938-1947 period and from the perspective of cancer, oil may be a more harmful fuel than coal, although coal is normally thought to be the more harmful fuel. Frankly, therefore, we could not offer an educated guess as to whether the 1938-1947 or 1950-1965 period would produce higher correlations.

Suffice it to say, the county unit is far from the best geographical unit to test this hypothesis. A better test would be the association between emissions of industrial sources, electricity generating, highway and other sources with census tract data. The census tracts could be combined to more accurately portray local terrain and meteorology than can much larger county units. Hopefully, in the future researchers in New Jersey will be able to make those finer scale geographical studies.

Summarizing, emissions from fossil fuel generating stations should positively associate with most of the body sites because some of the emissions are carcinogens and others can damage the lungs, digestive tract, and skin leading to neoplastic formations.

Motor Vehicles

Whereas fossil fuel generation data were readily available, data about motor vehicle travel varied by state. The following five data sets made the New Jersey information the most useful: (1) tabulations of accident factors that include not only accident data, but also lists of traffic lanes, traffic volumes and vehicle miles traveled (VMT)for every state road in every county; (2) summaries of available traffic data (for example, annual daily traffic) for specific control stations from 1948 to 1966 for all state roads and for major highways such as the New Jersey Turnpike and Garden State Parkway; (3) average annual daily traffic totals between specific milepost on the New Jersey Turnpike, the Garden State Parkway, and the Atlantic City Expressway from 1961-1975; (4) a list of total state highway VMT in New Jersey for every year from 1948-1975; and (5) state-wide traffic volume atlases.

These five data sets enabled us to construct an annual estimate of VMT for every county for the period 1945-1975. All of the data sets did not cover the same time periods. Accordingly, we verified the similarity of the four data sets for places for which comparable totals were available. Than we extrapolated to fill in missing data. A few illustrations will help explain the process.

After all the available state highway totals were calculated from the accident factor studies, the results were crosschecked with the state highway totals from the fourth source. The agreement was very close (e.g., 8,700,000,000 VMT in 1960 that we calculated and 8,418,000,000 from the DOT's list.)

For certain years no accident factor books exist, for example, 1945-1950. This problem was solved by applying the county total to the proportion of the closest year's actual percentage of VMT. If County X has 10 percent of the state VMT in 1951, and we did not have its 1950 total, we assigned it 10 percent of the state 1950 total.

The VMT's for the Garden State Parkway, New Jersey Turnpike and other major roads were calculated by using the VMT values between milepost markers. These roads cross county borders. Consequently, after the yearly VMT was calculated for each section of roadway, the state-wide traffic volume atlases had to be consulted in order to determine the percentage of the link that is in each county. The major road VMT values were added to the total VMT calculated from the state highway data to yield a year by year list of the VMT for each county of New Jersey from 1945 to 1975.

The data for Connecticut, Delaware, New York and Pennsylvania came from state and regional sources. Pennsylvania and Delaware VMT was supplied by the Pennsylvania Department of Transportation and the Delaware Valley Regional Planning Commission (DVRPC). The New York and Connecticut data are from the Tri-State Regional Planning Commission (TSRPC) and from state sources. The Pennsylvania data are in the form of annual VMT by county derived from Pennsylvania road log volumes. However, the data are only available for 1960, 1965, 1969, 1972-1973 and 1975. The New York and Connecticut data are also sparse; only 1963, and 1970-1974 were available.

Using New Jersey data, attempts were made to find an association between per capita VMT and demographic and economic factors. Had we been able to find a relationship it would have been used to simulate 1945, 1950, and 1955 values for the other states. In the absence of any relationship, the values were extrapolated.

Overall, we have annual VMT values for every county for the period 1945-1975. However, only the New Jersey data set is relatively reliable before 1960.

In light of the above data deficiencies, the VMT data were not transformed into direct measures of carcinogenic emissions such as benzo(a)pyrene. The weakness of the pre-1970-data outside of New Jersey is particularly unfortunate because post-1970 emission controls have greatly reduced benzo(a)pyrene emissions and a comparison would have been interesting. In addition, benzo(a)pyrene calculations might be interpreted as implying that benzo(a)pyrene is the only substance emitted by motor vehicles that is carcinogenic. There are others including asbestos from brakes and ozone which is probably a cocarcinogen, that is produced from the chemical reaction of partially burned fuel and sunlight. Asbestos from autos can constitute a hazard in places where many automobiles brake such as toll stations, tunnels, and auto repair shops. Whereas the automobile related asbestos hazard, as previously suggested, may be confined, ozone, an irritant that may cause a predisposition to lung cancer, may be spread far beyond the borders of the county where the emissions occur.

Despite data deficiencies and a lack of clarity about the role of motor vehicle emissions in carcinogenicity, we expected to find a positive correlation between VMT and lung and digestive tract cancers. We expected higher correlations by dividing VMT by land area, and still higher correlations from applying the distance-wind rose model to VMT data. As in the cases of particulates and electricity generating stations, the distance-wind rose model is speculative.

Water Supply and Soil

Without doubt, that most frustrating data problem we faced was finding surrogates for contaminated drinking water supplies and soils. Unfortunately, we never found satisfactory surrogates.

With respect to water supplies, we would have liked a historical record of samples for carcinogenic substances. Briefly, the available data consist of detailed records collected by the major public water companies. The records rarely include a lengthy list of carcinogenic substances, but are useful for the standard water quality parameters. The detailed records, however, have not been computerized or centralized in a single bureau. Therefore, they could not be collected for this phase of the study. We estimate that the study area has been 500 and 750 public potable water companies. Any attempt to gather the data for any but the largest of these will be a monumental task. The NJDEP's pioneering water sampling program now underway may provide useful information in the near future.

In the short run, one could gather data for the largest water purveyors and rely on the USEPA's computerized system (STORET) to provide information on the overwhelming majority of the region's water supplies. However, the author dismissed this possibility for this phase of the project for three reasons. First, the STORET data should be, but are not, screened for quality. As a result, bad data are mixed with accurate information to produce inaccurate data. Second, the summary printouts from STORET mix records from the 1940's and 1950's with records from the 1960's and 1970's. The results are averages that are relatively meaningless. Finally, whereas the quality of data in STORET leaves a good deal to be desired, the quantity is impressive. Having worked on a study of the Delaware estuary that used STORET data, the author estimates that requesting data covering the entire study area would have produced a printout between 3 and 5 feet high! In short, the use of STORET for studies like ours requires a capability to be more selective in data recovery than is presently possible.

To develop any water supply variables, we had to make some general assumptions about the relative utility of different sources of water. These are:

1. Surface water supplies carry more carcinogenic materials than groundwater supplies because they directly receive point discharges and runoff and are usually more heavily chlorinated;

2. Persons who use public potable water supplies are at higher risk than those who use their own personal wells because public potable water supplies are generally more polluted and are treated by chemicals.

Suffice it to say, these assumptions are general notions that will not hold in every place. Nevertheless, we expected that a positive association between surface water supplies and the digestive-urinary system cancers would appear. Conversely, a negative association between personal supplies and these cancers should appear.

With these hypotheses in mind, water supply data for the 45 areas in the study region were collected on the minor civil division level for the years 1963 and 1970. The intent was to determine the percentage of population served by public and private supplies of surface and groundwater. A good breakdown of this type of information was compiled by the Water Supply and Pollution Control Division of the U.S. Department of Health, Education, and Welfare in the 1963 Inventory - Municipal Facilities. Regional groupings of municipal water facilities provided an inventory of the population of the community, the estimated population served by the water facilities, the type of ownership, and the source or sources of water supply. When a population was served by a facility outside the municipal boundaries, the name of that facility was reported.

The information provided in the 1963 Inventory was essentially straightforward, with little problems encountered in data interpretation. The only adjustments made were for the facilities whose source of supply was both surface and groundwater. To determine the percentage of population served by these facilities, the proportion of surface water served of the total reporting either completely surface or ground was calculated for every county and applied to the municipalities reporting combined figures.

Water supply data for the 1970's, unfortunately, was not as easy to collect. First of all, there was no one source of information as compiled in the 1963 Inventory. This necessitated contacting each county planning board or regional commission for descriptive information delineating their sources of water supply. The information supplied by these requests varied in terms of usefulness. For the most part, the data were inconsistent. The majority of the county reports had supply figures for 1970; most of the remainder of the reports fluctuated between 1968 and 1969. The Pennsylvania counties in the study area reported the most detailed and most recent figures, with data from 1972 to 1975.

Primarily, the 1970 water supply data base was determined from the total average MGD yield from reporting facilities in each county. The total average yield from public and private facilities, by type of source, was then calculated as a percentage of the total average MGD yield for the county as a whole. This was accomplished for all but fifteen of the 45 counties in the study area. Of these fifteen, five counties in New York provided maps of supply facility locations that neglected to indicate the estimated population served or the type of facility ownership. The 10 remaining counties failed to provide any information on their water supply facilities. With data on one-third of the places missing, we had to discontinue the analysis of the 1970 data.

The final variables are as follows:

1. % of county population served by public potable surface water supplies, 1963;
2. % of county population served by public potable, groundwater supplies, 1963;
3. % of county population served by public potable supplies, 1963; and
4. % of county population served by their own supplies, 1963.

No variables were defined for soils. We would have liked to have had indicators of the metal and chlorinated hydrocarbon content of soils, but no data were available.

Summary

This chapter has described the cancer mortality data and how high and low risk factors drawn from the literature have been developed in order to study the geography of cancer mortality and to perform a statistical analysis of the relationship between the geographical distribution of cancer and of risk factors. The study area includes 49 counties that roughly correspond to the New Jersey-New York City-Philadelphia metropolitan regions. The cancer data set is the National Cancer Institute's landmark 1950-1969 county tabulations. Regression, correlation and factor analysis are used to study the association between cancer mortality and high and low risk factors.

The use of these and other statistical techniques may lead to false positive, false negative, and confusing results. The causes of these problems are reviewed and guidelines for controlling or at least minimizing spurious results are considered. These include drawing hypotheses from the literature, testing the results for consistency and strength, using control diseases and control independent variables, and in general, recognizing the limitations of mathematical tools.

Socioeconomic status, ethnicity, stability, and fertility data were gathered to represent high risk personal environment factors. Missing information include patterns of smoking, drinking, drug consumption and other risk factors.

Occupational data at the 2-digit standard industrial code level were available by place of work and residence. The occupational data were converted into a variety of indices of economic specialization.

Local ambient environment data sets were the poorest. Air quality and emissions data were combined, but particularly the air quality data are questionable. A simple model was developed to study the diffusion of emissions and of fossil fuel and motor vehicle data. Finally, limited tabulations of each county's types of water service were available.

In the next chapter, we present an analysis of the spatial distribution of cancer mortality in the study area. Chapters 3 and 4 present the associations between cancer mortality and high risk factors.

Before leaving the data chapter, it is important to once again emphasize the limitations placed on the researcher by the quality of the data. When combined with the large size of the units of analysis, the poor quality of some of the data that we were forced to use implies that the most appropriate function of the study is to isolate disease-place combinations for further study by persons who will have access to better data.

CHAPTER 2

SPATIAL ASSOCIATION OF CANCER MORTALITIES IN THE NEW JERSEY-NEW YORK-PHILADELPHIA METROPOLITAN AREAS

This chapter presents the association between mortality recorded for thirteen cancer body sites among four population groups in a forty-nine county study area. The chapter reports a definite tendency for groups of spatially contiguous counties to strongly manifest particular cancer mortalities. These geographical patterns in turn, suggest associations with urbanization, industrialization, air quality, income, national origin, and other factors hypothesized to be involved with variation in cancer rates. While some of these geographical patterns may be striking and the obvious ones are noted, we have refrained from explicity reviewing any testing of causality in this chapter. Chapters 3 and 4 will explicitly test and report on the geographical association among the cancer mortality rates and the high and low risk factors.

This chapter is divided into three parts: (1) selection of data; (2) comparison of the region and the nation; and (3) description of intraregional variation.

Selection of Data

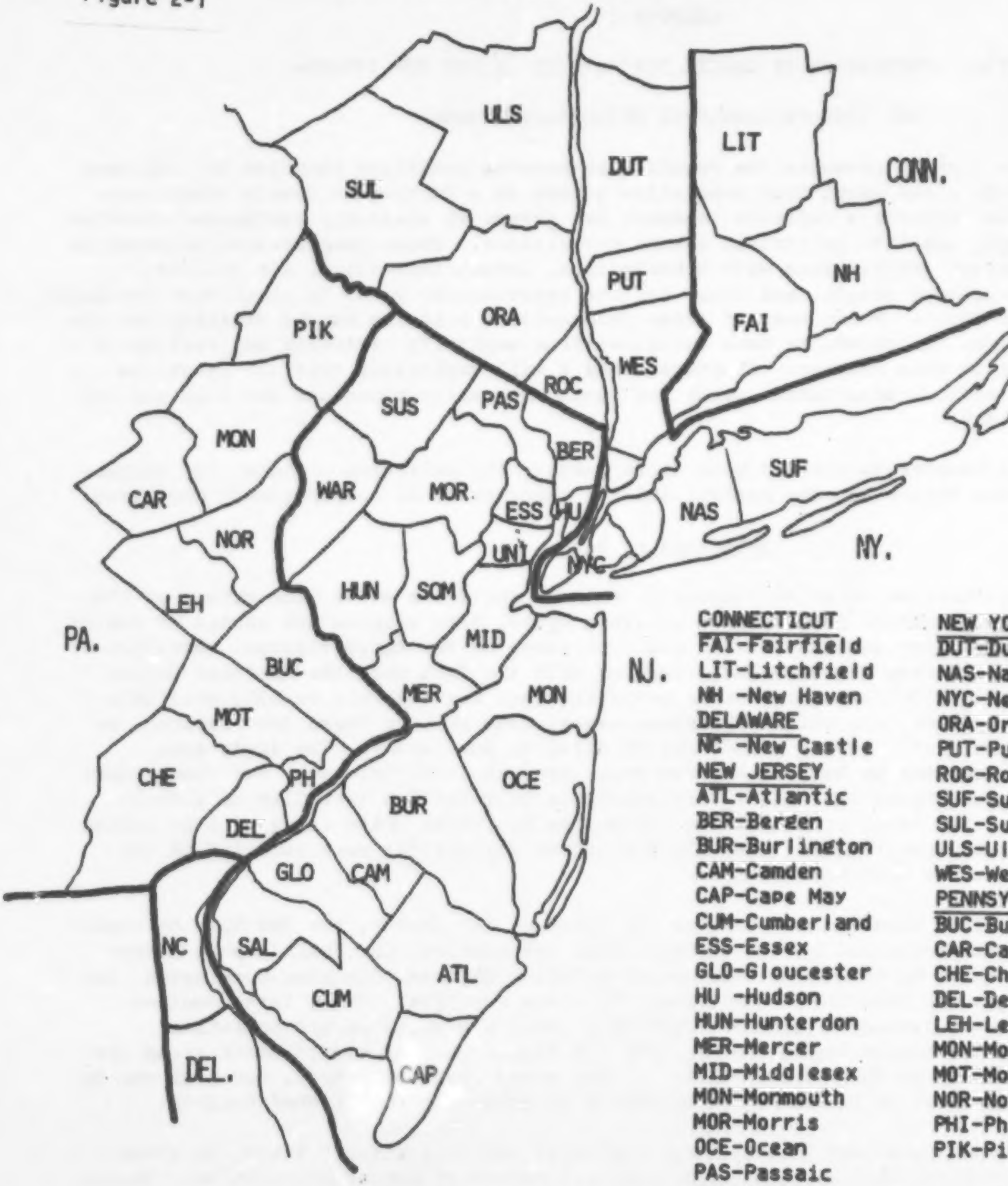
Initially, as noted in chapter 1, we made decisions about the extent of the study area, scale of geographical analysis units, data sources and choice of cancer sites. The major interest of the research group is the intra-regional variation of cancer in an urban region. When combined with the fact that the National Cancer Institute's 1950-1969 county scale mortality data was the only readily available set of age, race, and sex standardized cancer data when we began the research, we chose a forty-nine county study area centered on New Jersey. The study area includes counties in New Jersey, New York, Pennsylvania, Delaware, and Connecticut (Table 9 and Figure 1). Rates were available only for New York City as a whole. Therefore, the study area includes forty-five political units (they will be called areas in the study). More than 850,000 cancer mortalities were recorded in the study area from 1950 through 1969.

The study area includes all of the State of New Jersey, the New York Metropolitan Region as defined by the Regional Plan Association, the Philadelphia Metropolitan Region as defined by the Delaware Valley Regional Planning Commission, and some additional counties bordering on the above counties. These last counties were selected because we assumed that they contain sources of air pollution impacting the larger region and/or they are second home or resettlement areas for persons living in the central cities of the study area. In short, our goal was to maintain as much as possible the integrity of commuting and airshed regions.

Two considerations governed the choice of the body sites. First, we chose those body sites that represent the greatest number of cancer mortalities. Second, body sites that the literature suggests might be related to the physical and social environments found in the study area were selected. One reviewer felt that more body sites should have been analyzed (e.g., cervix, esophagus). More sites were not included because some of the statistical methods that have been used tend to produce gibberish results when the number of diseases is large relative to the number of geographical areas. Several reviewers felt that too many diseases were analyzed. In retrospect, given the weak results for some of the sites, perhaps too many diseases were included. Having performed the analyses, we will report not only the strong but also the weak findings.

Figure 2-1

FORTY-NINE COUNTY STUDY AREA


ULS
LIT
DUT
CONN.
SUL
PUT
NH
ORA
PIK
FAI
WES
ROC
PAS
SUS
MON
BER
SUF
WAR
MOR
CAR
ESS
HU
NAS
NOR
UNI
NYC
NY.
HUN
SOM
LEH
MID
PA.
BUC
NJ.
MER
MON
MOT
OCE
CHE
PH
BUR
DEL
GLO
CAM
NC
SAL
ATL
CUM
DEL.
CAP

CONNECTICUT
FAI-Fairfield
LIT-Litchfield
NH -New Haven
DELAWARE
NC -New Castle
NEW JERSEY
ATL-Atlantic
BER-Bergen
BUR-Burlington
CAM-Camden
CAP-Cape May
CUM-Cumberland
ESS-Essex
GLO-Gloucester
HU -Hudson
HUN-Hunterdon
MER-Mercer
MID-Middlesex
MON-Monmouth
MOR-Morris
OCE-Ocean
PAS-Passaic
SAL-Salem
SOM-Somerset
SUS-Sussex
UNI-Union
WAR-Warren

NEW YORK
DUT-Dutchess
NAS-Nassau
NYC-New York City
ORA-Orange
PUT-Putnam
ROC-Rockland
SUF-Suffolk
SUL-Sullivan
ULS-Ulster
WES-Westchester
PENNSYLVANIA
BUC-Bucks
CAR-Carbon
CHE-Chester
DEL-Delaware
LEH-Lehigh
MON-Monroe
MOT-Montgomery
NOR-Northampton
PHI-Philedelphia
PIK-Pike

TABLE 9

Forty-Nine County Study Area

1. Atlantic, New Jersey
2. Bergen, New Jersey
3. Burlington, New Jersey
4. Camden, New Jersey
5. Cape May, New Jersey
6. Cumberland, New Jersey
7. Essex, New Jersey
8. Gloucester, New Jersey
9. Hudson, New Jersey
10. Hunterdon, New Jersey
11. Mercer, New Jersey
12. Middlesex, New Jersey
13. Monmouth, New Jersey
14. Morris, New Jersey
15. Ocean, New Jersey
16. Passaic, New Jersey
17. Salem, New Jersey
18. Somerset, New Jersey
19. Sussex, New Jersey
20. Union, New Jersey
21. Warren, New Jersey
22. Fairfield, Connecticut
23. Litchfield, Connecticut
24. New Haven, Connecticut
25. Dutchess, New York
26. Nassau, New York
27. New York City, New York*
28. Orange, New York
29. Putnam, New York
30. Rockland, New York
31. Suffolk, New York
32. Sullivan, New York
33. Ulster, New York
34. Westchester, New York
35. Bucks, Pennsylvania
36. Carbon, Pennsylvania
37. Chester, Pennsylvania
38. Delaware, Pennsylvania
39. Lehigh, Pennsylvania
40. Monroe, Pennsylvania
41. Montgomery, Pennsylvania
42. Northampton, Pennsylvania
43. Philadelphia, Pennsylvania
44. Pike, Pennsylvania
45. New Castle, Delaware

*Aggregation of five-county area in the mortality data

The twelve specific body sites (all sites is the thirteenth) account for the vast majority of cancer mortalities among the white and non-white male and female populations (Table 10). Specifically, the twelve sites account for about 75 percent of all the male mortalities and between 55 and 66 percent of the female cancer mortalities.

TABLE 10

Cancer Mortality By Body Site Per 100,000
National and Regional Averages, 1950-1969

	REGION				U.S.A.			
BODY SITE	WHITE		NON-WHITE		WHITE		NON-WHITE	
	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE	MALE	FEMALE
1. Stomach	17.09*	8.72	25.95*	10.23	15.22	7.70	24.03	10.69
2. Large Intestine	20.34**	20.10**	14.91	16.10	16.54	16.25	12.07	12.69
3. Rectum	10.42**	6.16*	7.73	5.67	7.65	4.82	5.68	4.46
4. Liver[a]	5.19	5.52	6.57	5.94	5.16	5.34	6.91	4.59
5. Pancreas	10.01	6.21	11.69	6.52	9.63	5.83	10.17	6.17
6. Lung[b]	42.68*	6.93*	51.76*	8.03*	37.98	6.29	36.67	6.27
7. Breast	___	28.83**	___	24.05	___	25.51	___	22.10
8. Prostate	17.72	___	29.59	___	17.84	___	27.39	___
9. Kidney	4.24	2.07	3.60	1.51	3.86	1.99	2.67	1.42
10. Bladder	9.24**	2.83	6.63*	3.45	6.78	2.39	5.05	3.05
11. Skin[c]	2.91	1.72	1.53	0.86	3.06	1.82	1.25	0.92
12. Leukemia	8.56	5.69	6.85	4.25	8.81	5.74	5.77	3.88
TOTAL 1-12	148.40	94.78	166.81	86.61	132.53	83.68	137.66	76.24
% of all sites	77	66	74	57	76	64	75	55
13. All Sites	192.12**	142.78**	223.91*	153.03	174.04	130.10	184.28	139.18

*Exceeds 80th percentile for all states
**Exceeds upper decile for all states.

a Includes Biliary Passages
b Includes Trachea and Bronchus
c Melanoma and other skin

Cancer Mortality: Comparison of the Region and the Nation

While this study focuses on intraregional variation in cancer mortality, regional and national rates will be compared to provide an initial perspective. A glance at Table 10 will suggest that most of the region's rates are higher than the national averages. A better perspective is provided by Table 11. The regional averages have been divided by the national averages to obtain 48 ratios.

The region's high cancer mortality rates are evident. Forty of the 48 ratios exceed unity. In the case of male rectum and bladder cancer, the regional average for whites and nonwhites is more than thirty percent above the national average. Four of the twelve nonwhite male rates exceed the national average for nonwhite males by more than thirty percent. The total nonwhite male cancer mortality rate exceeds the national norm for the nonwhite population by twenty-two percent. While the white male population exceeds the national average for the male white population by a smaller proportion than the nonwhite male population exceeds its comparative group, the white male regional average exceeds the upper decile for the fifty states.

In short, during 1950-1969 the study area's male population had considerably higher cancer mortality rates than the male population of the United States as a whole. The same generalization is true to a lesser extent of the region's female population. Overall when controlled for age differences, the region had cancer mortality rates ten percent higher than the nation's rates.

In light of recent statements that New Jersey no longer has the highest cancer mortality rates in the United States, it would be useful to present a comparison of the region's and the nation's recent cancer mortality experiences and changes in cancer rates from the 1950's through the 1970's. We have made such comparisons, and they will be presented in Parts II and III of this study.

TABLE 11

Cancer Mortality by Body Site:
Ratio of Regional to National Average, 1950-69

		WHITE REGION/US AVERAGE		NONWHITE REGION/US AVERAGE	
	BODY SITE	MALE	FEMALE	MALE	FEMALE
1.	Stomach	1.12*	1.13	1.08*	0.96
2.	Large Intestine	1.23**	1.24**	1.24	1.27
3.	Rectum	1.36**	1.28*	1.36	1.27
4.	Liver	1.01	1.03	0.95	1.29
5.	Pancreas	1.04	1.07	1.15	1.06
6.	Lung	1.12*	1.10*	1.41*	1.28*
7.	Breast	—	1.13**	—	1.09
8.	Prostate	0.99	—	1.08	—
9.	Kidney	1.10	1.04	1.35	1.06
10.	Bladder	1.36**	1.18	1.31*	1.13
11.	Skin	0.95	0.95	1.22	0.93
12.	Leukemia	0.97	0.99	1.19	1.10
13.	All	1.10**	1.10**	1.22*	1.10

*Exceeds 80th percentile for all states.
**Exceeds upper decile for all states.

Description of Intraregional Variation

Maps, bivariate, and multivariate analyses were used to isolate intraregional differences. Maps were used as an initial screening device and later to summarize the multivariate analyses. Bivariate correlations indicate the statistical covariation between pairs of cancer mortalities. Multivariate analysis shows the coincidence between two or more cancer mortalities.

Patterns Revealed by Mapping Diseases

We mapped the cancer mortalities stressing intraregional differences at the county scale rather than depicting each county's rates relative to the national average, decile, or other level. This decision reflects our interest in intraregional differences. Second, it reflects the fact that maps of cancer mortality in our study are keyed to national rates in many cases are not revealing because so many of the county rates are considerably above national averages that a preponderance of the cases fall into one category. For example, the national average and upper decile for the white male population are 174.04 and 178.06, respectively. An intraregional map based on those two numbers would not be very revealing. It would have three counties in a less than national average category, four in a category between the national average and upper decile for the states, and thirty-eight in the greater than upper national decile category.

In contrast to mapping on the basis of national rates, we prepared maps with approximately the upper quartile of counties in one group, the lower quartile in a second group, and the middle fifty percent in a third group. This 25-25-50 working rule was modified when obvious breaks occurred in the data.

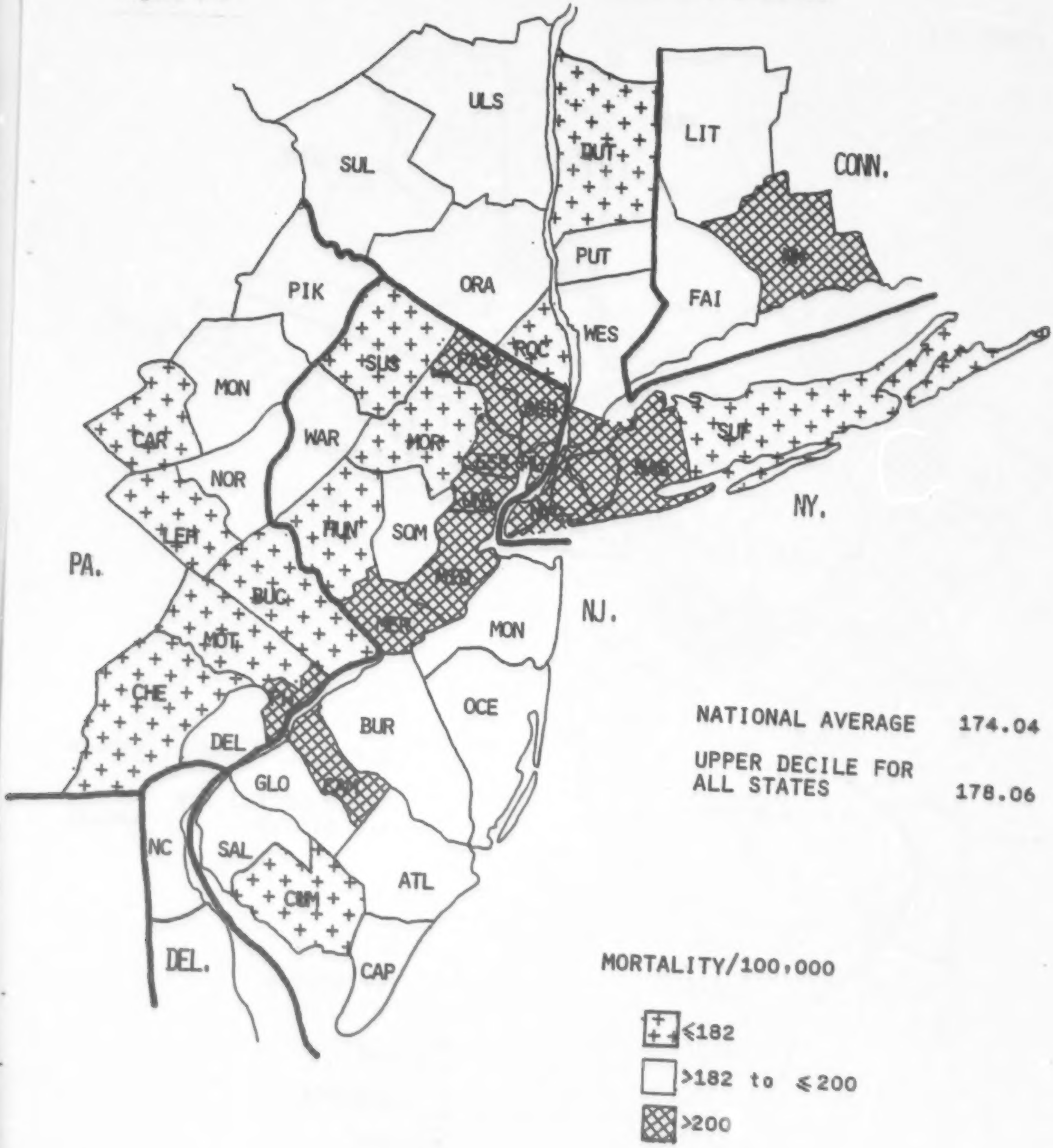
The male white all cancer mortality map appears below (Figure 2). From a technical perspective, Figure 2 is a much more revealing map than one which would have had the overwhelming majority of the cases in one category. The pattern portrayed by Figure 2 is striking. The twelve counties with mortality rates fifteen percent or more above the national average (greater than 200/100,000) lie along the urban-industrial corridor; the twelve counties with rates closer to the national average (less than or equal to 182/100,0000 lie in less developed portions of the region.

Before concluding that living along the region's urban-industrial spine leads to higher cancer mortality rates, one would have to test the coincidence among indicators of an urban-industrial environment and cancer mortality through painstaking epidemiological research. While waiting for such evidence to be developed, we can look for indirect evidence by comparing the male all cancer mortality map with a similar map for the female and nonwhite populations of the region. If causal factors impacting on the urban-industrial corridor are extremely strong, they should reappear among all the four population groups. Figures 3-5 are these maps.

A strong correspondence appears among the male and female white maps. Nine of twelve political units in the highest category in the male white map also are in the highest mortality category in the female population map: Nassau, New York City, Bergen, Essex, Hudson, Union, Middlesex, Philadelphia and Camden. However,

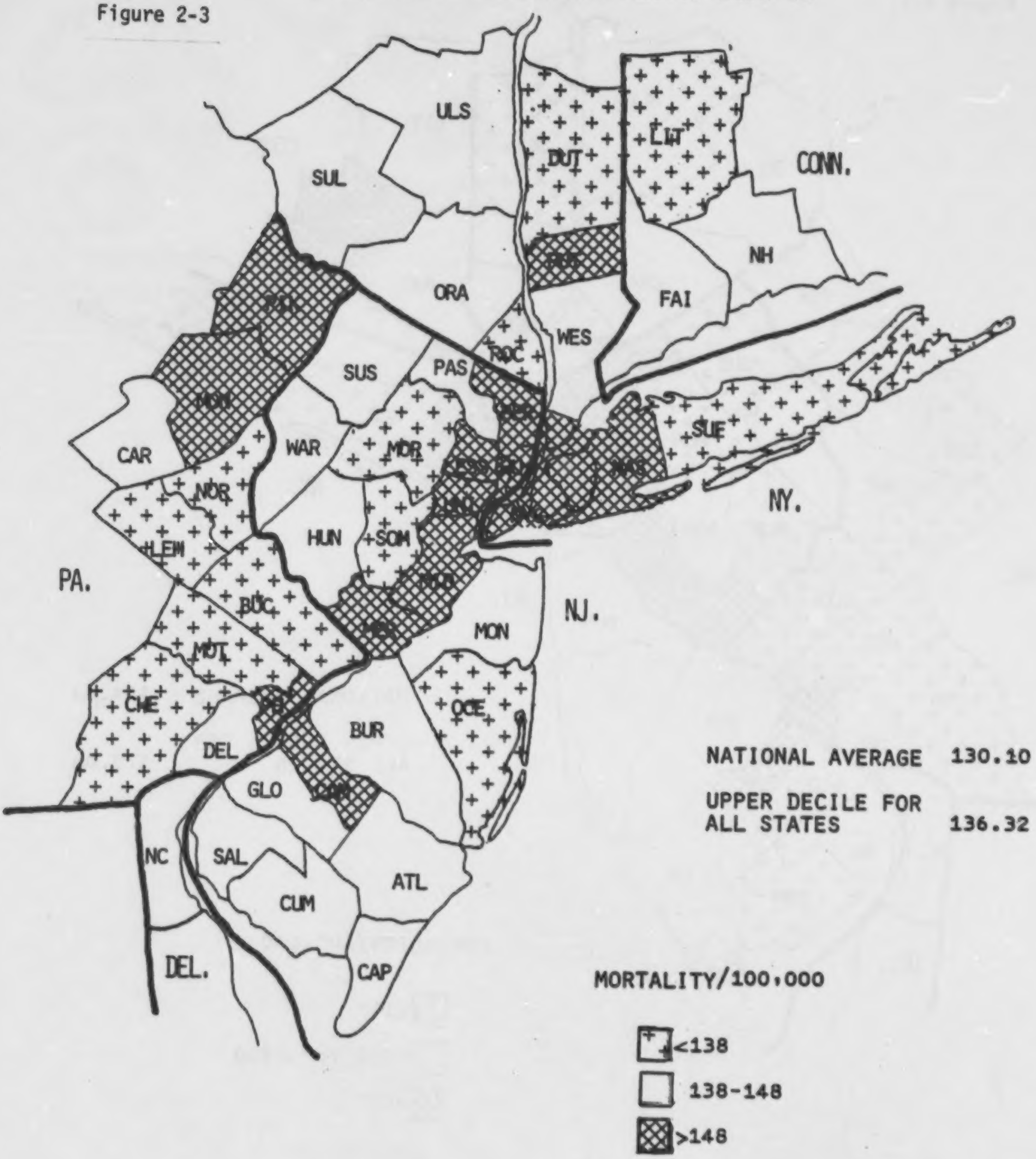
Figure 2-2

MALE WHITE CANCER MORTALITY, 1950-69


ULS
SUL
DUT
LIT
CONN.
PUT
PIK
ORA
FAI
WES
ROC
SUS
MON
CAR
WAR
MOR
SUF
NOR
LEH
HUN
SOM
NY.
PA.
BUC
NJ.
MON
MOT
CHE
OCE
BUR
DEL
GLO
NC
SAL
ATL
CUM
DEL.
CAP
NATIONAL AVERAGE 174.04
UPPER DECILE FOR ALL STATES 178.06
MORTALITY/100,000
≤182
>182 to ≤200
>200

FEMALE WHITE CANCER MORTALITY, 1950-69

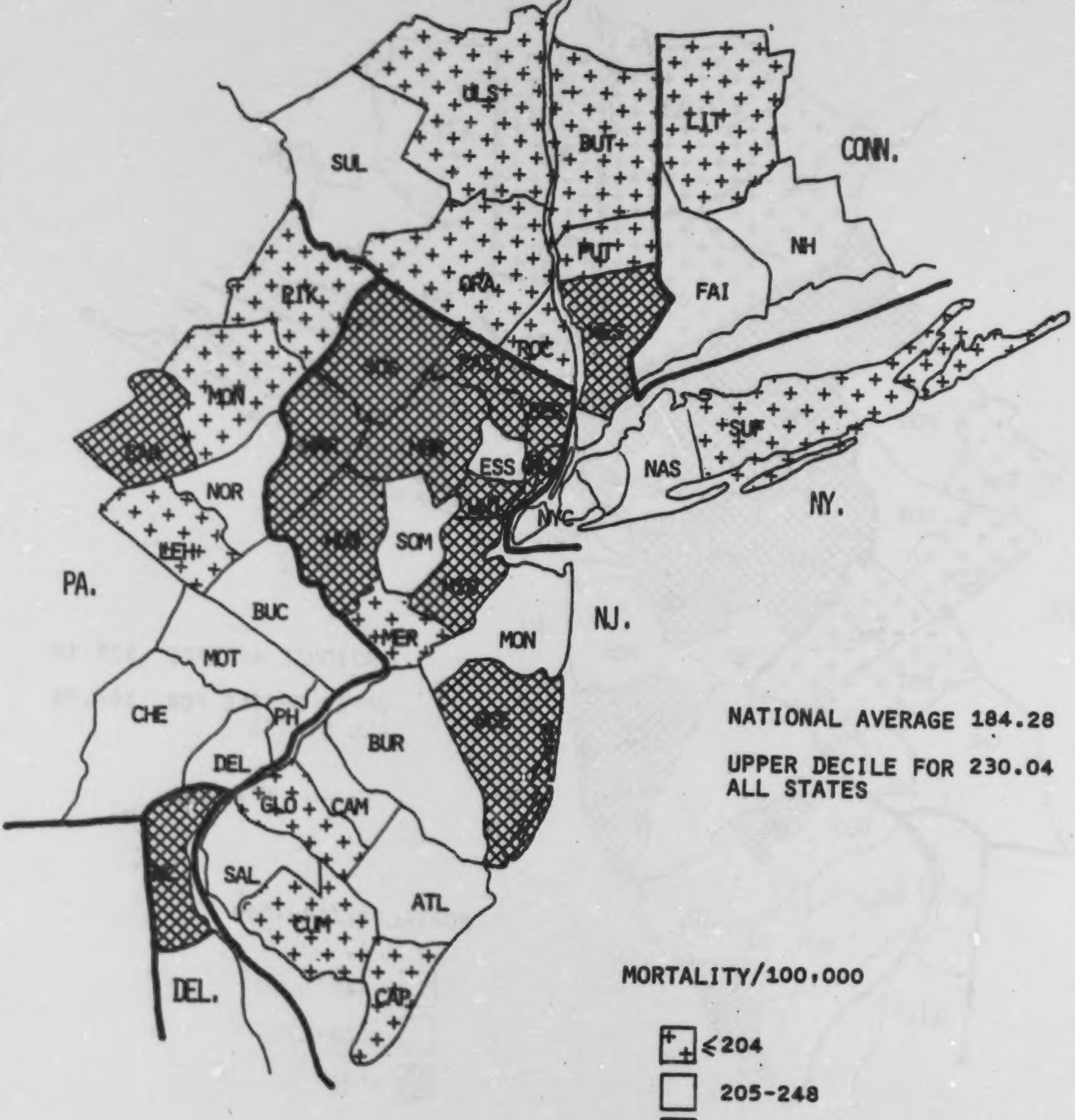
Figure 2-3


ULS
SUL
DUT
LIT
CONN.
NH
FAI
ORA
ROC
WES
PAS
SUS
SUF
CAR
WAR
MOR
NOR
NY.
LEH
HUN
SOM
PA.
BUC
NJ.
MON
MOT
CHE
OCE
BUR
DEL
GLO
NC
SAL
ATL
CUM
DEL.
CAP
MORTALITY/100,000
<138
138-148
>148

NATIONAL AVERAGE 130.10

UPPER DECILE FOR ALL STATES 136.32

Figure 2-4 MALE NONWHITE CANCER MORTALITY, 1950-69


ULS
BUT
LIT
CONN.
SUL
NH
PUT
FAI
RIK
ORA
ROC
MON
SUF
NOR
ESS
NAS
NY.
NYC
LEH
SOM
PA.
BUC
MER
NJ.
MON
MOT
CHE
PH
BUR
DEL
GLO
CAM
SAL
ATL
CUM
CAP
DEL.

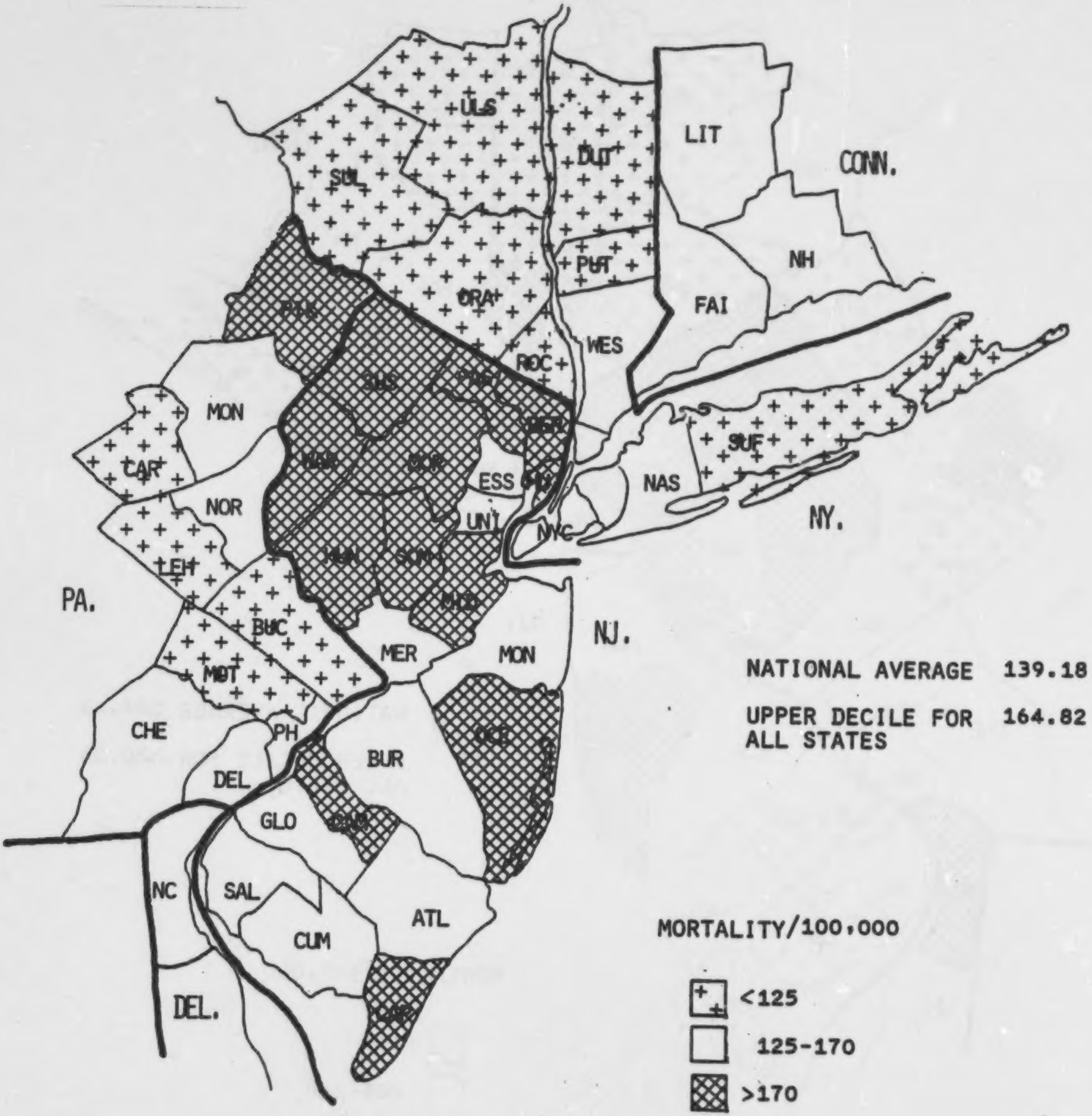
NATIONAL AVERAGE 184.28

UPPER DECILE FOR 230.04
ALL STATES

MORTALITY/100,000

- + ≤204
- ☐ 205-248
- ▩ >248

Figure 2-5 FEMALE NONWHITE CANCER MORTALITY, 1950-69


ULS
DUT
LIT
CONN.
SUL
PUT
NH
FAI
ORA
WES
ROC
SUS
MON
CAR
WAR
MOR
ESS
NAS
SUF
NOR
UNI
NYC
NY.
LEH
HUN
SOM
MID
PA.
BUC
NJ.
MER
MON
MGT
CHE
PH
BUR
OCE
DEL
GLO
CAM
NC
SAL
ATL
CUM
CAP
DEL.
MORTALITY/100,000
<125
125-170
>170

NATIONAL AVERAGE 139.18

UPPER DECILE FOR ALL STATES 164.82

relatively rural Pike, Putnam, and Monroe Counties, contrary to the urban-industrial pattern, show high female white cancer mortality rates. Several explanations may be offered to explain the high rates of these three counties. First, the major causes of cancer which affect white females (breast, uteri) are not necessarily related to urban-industrial living. Second, the excess mortalities in these counties might reflect migration to these counties from the urban-industrial corridor. (Part II of this study will discuss this factor in some detail.) The above average mortalities for the region might also reflect the local environment, ethnic, family and other factors. Finally, since these three counties combined have had less than 1,600 cancer-related female mortalities (compared to almost 140,000 for New York City), it could be argued that the relatively few extra cases that caused these high mortality rates are due to intraregional variations in the recording of the causes of death. Overall, the male and female white mortality patterns are similar, but far from identical.

The nonwhite population's cancer mortality rates are higher than the white population's. Accordingly, one would expect an even more marked concentration of cancer mortalities in the urban-industrial corridor region. Figures 4 and 5 portray the expected pattern to some extent, but the striking impression of the map is that nonwhites in northern New Jersey have high cancer mortality rates. Densely developed New York City and Philadelphia inexplicably are not included in the highest group, while rural Sussex, Warren and Hunterdon Counties in New Jersey have extremely high rates. It is probable that the high rates in the rural New Jersey counties are artifacts. Specifically, with respect to the female nonwhite population, Pike, Sussex, Warren and Hunterdon Counties had only fifty-three cancer mortality cases. A small reduction in the number of cases or systematic difference in diagnosis could account for the high rates. On the other hand, the geographical contiguity of these counties and the fact that they border most of the other high nonwhite mortality counties, suggests that they should not be so readily explained away as artifacts. Summarizing, at the county scale there seems to be an association between total male and female white mortality rates and a lesser association between white and nonwhite total cancer mortality rates.

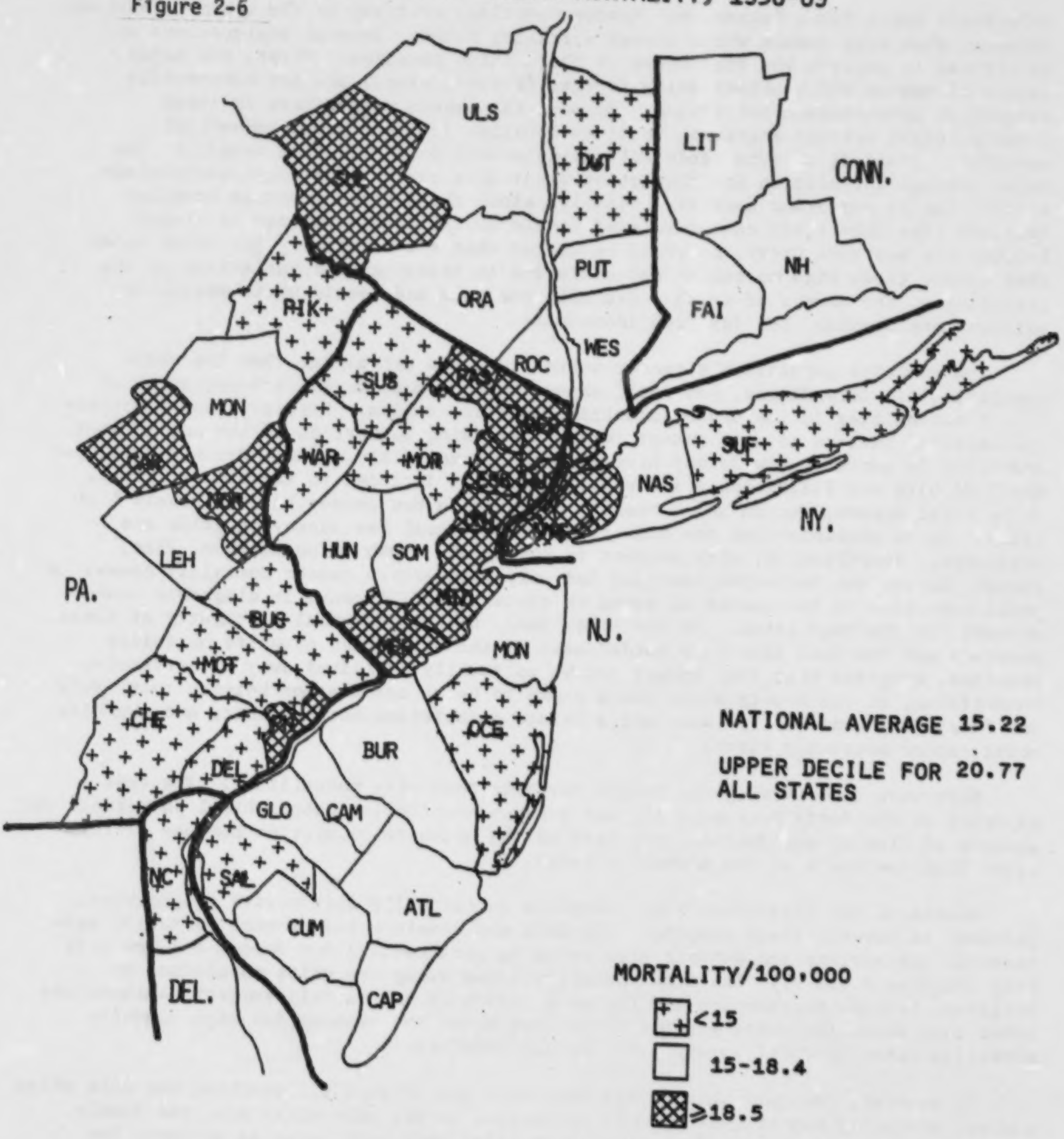
Maps were also made of the twelve specific body site mortalities. The vast majority of the forty-four maps did not produce distinctive geographical patterns. The absence of clarity was particularly true of the nonwhite population because so few cases form the basis of the mortality rates.

Eleven of the forty-four maps exhibited sufficiently distinctive geographical patterns to warrant final mapping. The male and female white stomach mortality maps resemble one another and exhibit high rates in northeastern New Jersey and New York City (Figures 6 and 7). The high mortality rates among the white population in Sullivan, Carbon, Northampton and Putnam Counties do not readily suggest explanations other than those that were previously offered about the reasons for high nonwhite mortality rates in rural northern New Jersey Counties.

In general, the male white large intestine map (Figure 8) resembles the male white stomach mortality map (Figure 6). In comparison to the male white map, the female white large intestine map (Figure 9) shows relatively high rates in southern New Jersey.

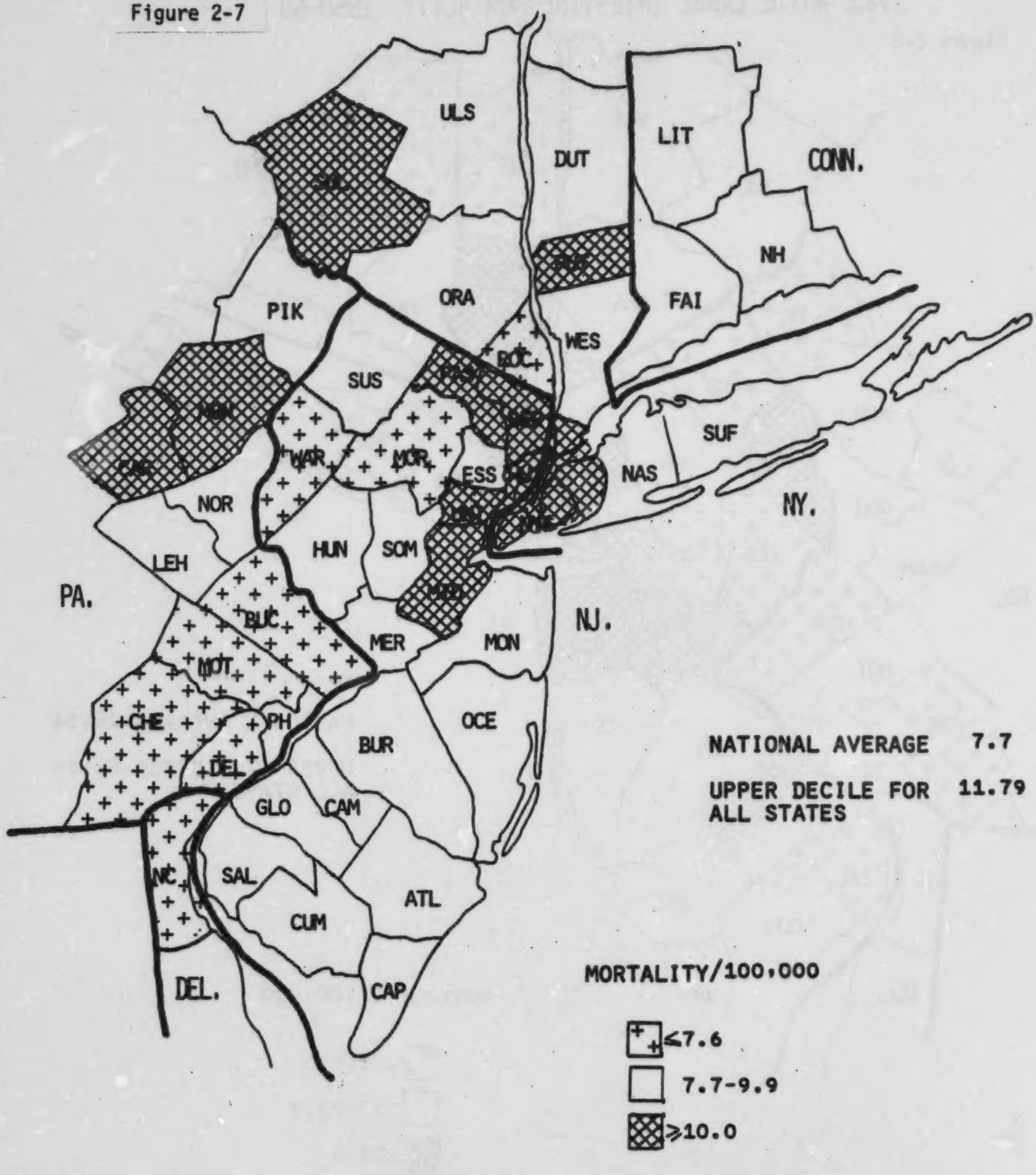
MALE WHITE STOMACH CANCER MORTALITY, 1950-69

Figure 2-6


ULS
DUT
LIT
CONN.
NH
PUT
PIK
ORA
FAI
ROC
WES
SUB
MON
SUF
WAR
MOR
NAS
NY.
LEH
HUN
SOM
PA.
BUS
NJ.
MON
MOT
CHE
OCE
BUR
DEL
GLO
CAM
NC
SAL
ATL
CUM
DEL.
CAP
NATIONAL AVERAGE 15.22
UPPER DECILE FOR ALL STATES 20.77
MORTALITY/100,000
<15
15-18.4
≥18.5

FEMALE WHITE STOMACH CANCER MORTALITY, 1950-69

Figure 2-7

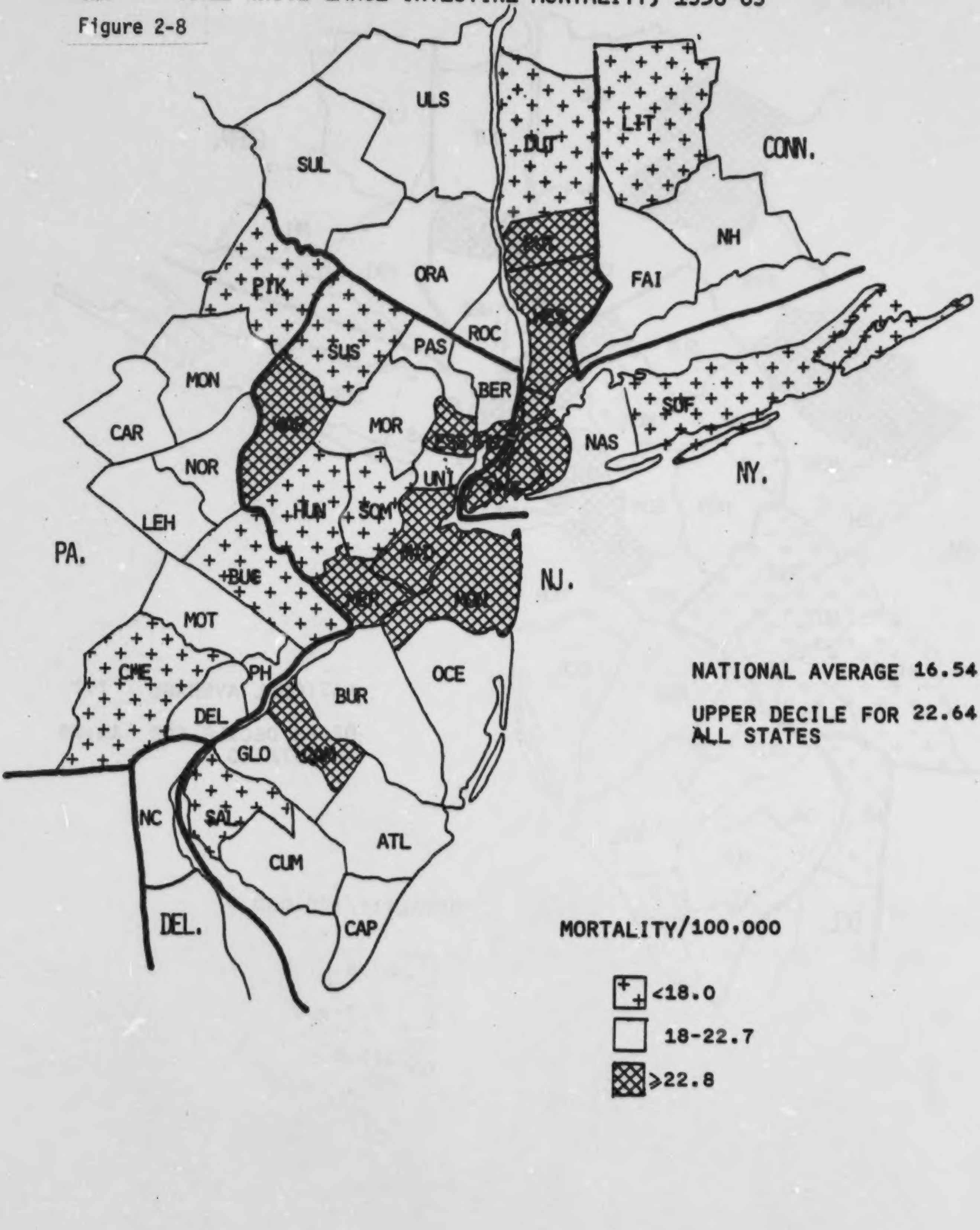

ULS
DUT
LIT
CONN.
NH
FAI
ORA
PIK
ROC
WES
SUS
SUF
NAS
NY.
WAR
MOR
ESS
NOR
LEH
HUN
SOM
PA.
BUC
MER
MON
NJ.
MOT
CHE
PH
OCE
BUR
DEL
GLO
CAM
NC
SAL
ATL
CUM
DEL.
CAP
MORTALITY/100,000
≤7.6
7.7-9.9
≥10.0

NATIONAL AVERAGE 7.7

UPPER DECILE FOR ALL STATES 11.79

MALE WHITE LARGE INTESTINE MORTALITY, 1950-69

Figure 2-8



ULS
SUL
DUT
LIT
CONN.
NH
PIK
ORA
FAI
ROC
PAS
SUS
BER
SUF
MON
MOR
CAR
NAS
NOR
UNI
NY.
HUN
SOM
LEH
PA.
BUC
N.J.
MOT
MER
MON
CHE
PH
OCE
BUR
DEL
GLO
CAM
NC
SAL
ATL
CUM
DEL.
CAP

NATIONAL AVERAGE 16.54

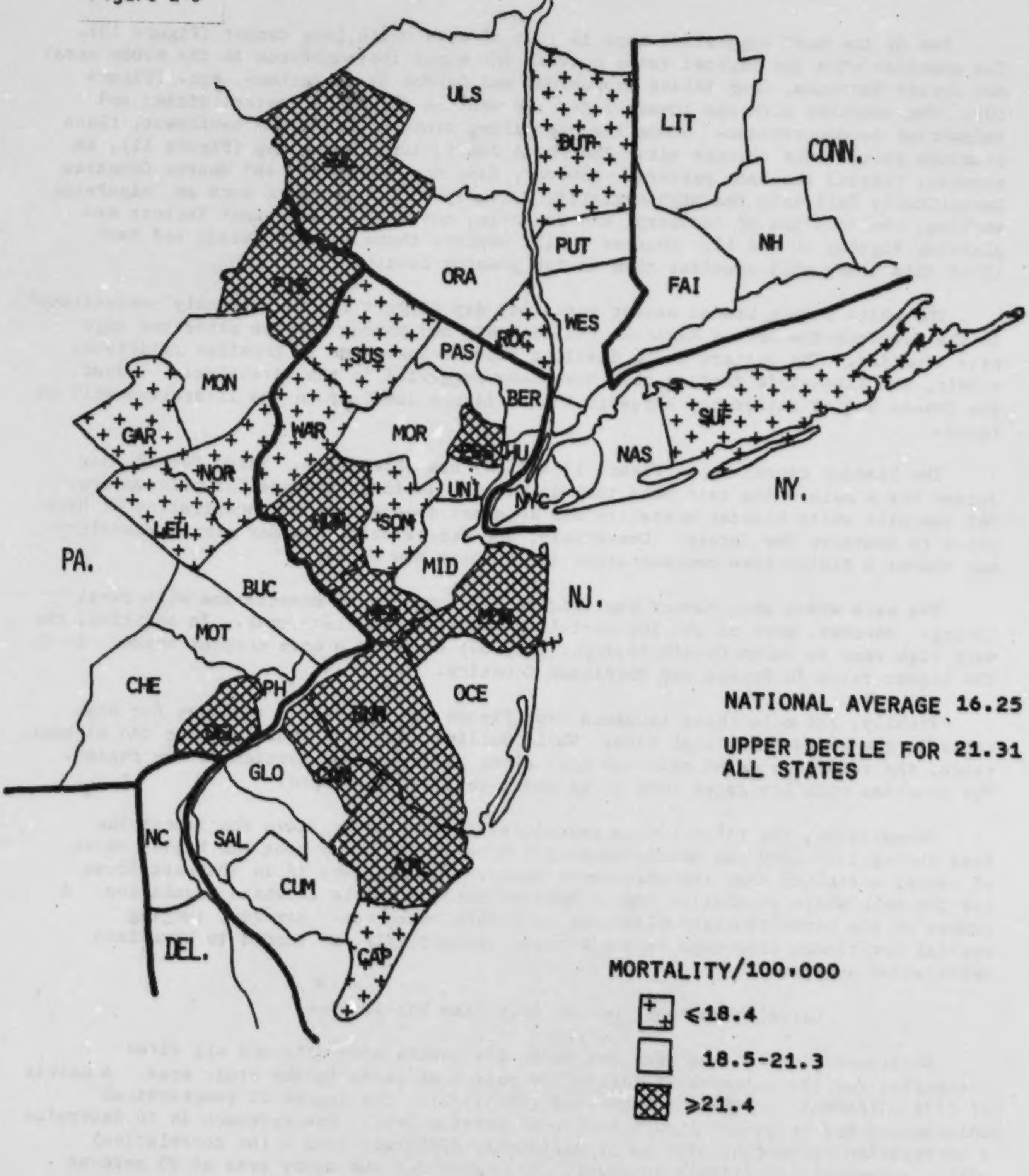
UPPER DECILE FOR ALL STATES 22.64

MORTALITY/100,000

+ <18.0
☐ 18-22.7
▨ ≥22.8

FEMALE WHITE LARGE INTESTINE MORTALITY, 1950-69

Figure 2-9


ULS
LIT
CONN.
DUT
SUL
PUT
NH
ORA
FAI
PIK
WES
ROC
SUS
PAS
MON
BER
SUF
GAR
WAR
MOR
ESS
HU
NAS
NOR
UNI
NYC
NY.
LEH
HUN
SOM
PA.
MID
BUC
NJ.
MER
MON
MOT
CHE
PH
OCE
BUR
DEL
GLO
CAM
NC
SAL
ATL
CUM
DEL.
CAP
MORTALITY/100,000
≤18.4
18.5-21.3
≥21.4

NATIONAL AVERAGE 16.25

UPPER DECILE FOR ALL STATES 21.31

One of the most suggestive maps is that of male white lung cancer (Figure 10). The counties with the highest rates contain the major thoroughfares in the study area: New Jersey Turnpike, Long Island Expressway and Garden State Parkway, etc. (Figure 10). The counties with the lowest rates lie west of the major central cities and industrial concentrations. Since the prevailing winds are from the southwest, these counties should have cleaner air. The white female lung cancer map (Figure 11), in general, retains the same pattern. However, Cape May, Atlantic, and Monroe Counties inexplicably fall into the high mortality category. Other factors such as cigarette smoking, the location of industry, and commuting may also be important factors explaining Figures 10 and 11. Chapter 4 will explore those in more detail and Part II of this study will consider them in far greater detail.

The white female breast cancer mortality map (Figure 12) is extremely concentrated in the New York-New Jersey Region. Philadelphia and Monroe are the other two high rate counties. The pattern is so distinct that it leads one to consider religious, ethnic, and life-style factors that have been suggested in the literature. Indeed, the female breast cancer map suggests that evidence contrary to the literature will be found.

The bladder cancer maps (Figure 13 and 14) are intriguing. Salem County, New Jersey has a male white rate more than double the national average for male whites. Yet the male white bladder mortality map does not demonstrate a concentration of high rates in southern New Jersey. Conversely, the female white bladder cancer mortality map showed a distinctive concentration in southern New Jersey.

The male white skin cancer map (Figure 15) suggests an association with rural living. However, most of the low mortality counties are also rural. In addition, the very high rate in Salem County (4.7 per 100,000) suggests a more complex answer, as do the higher rates in Nassau and Middlesex Counties.

Finally, the male white leukemia map (Figure 16) suggests a tendency for high rates in the urban-industrial core. While Sullivan and Pike Counties have the highest rates, the remaining seven counties fall along the urbanized portion of the region. The counties with low rates tend to be among the less developed.

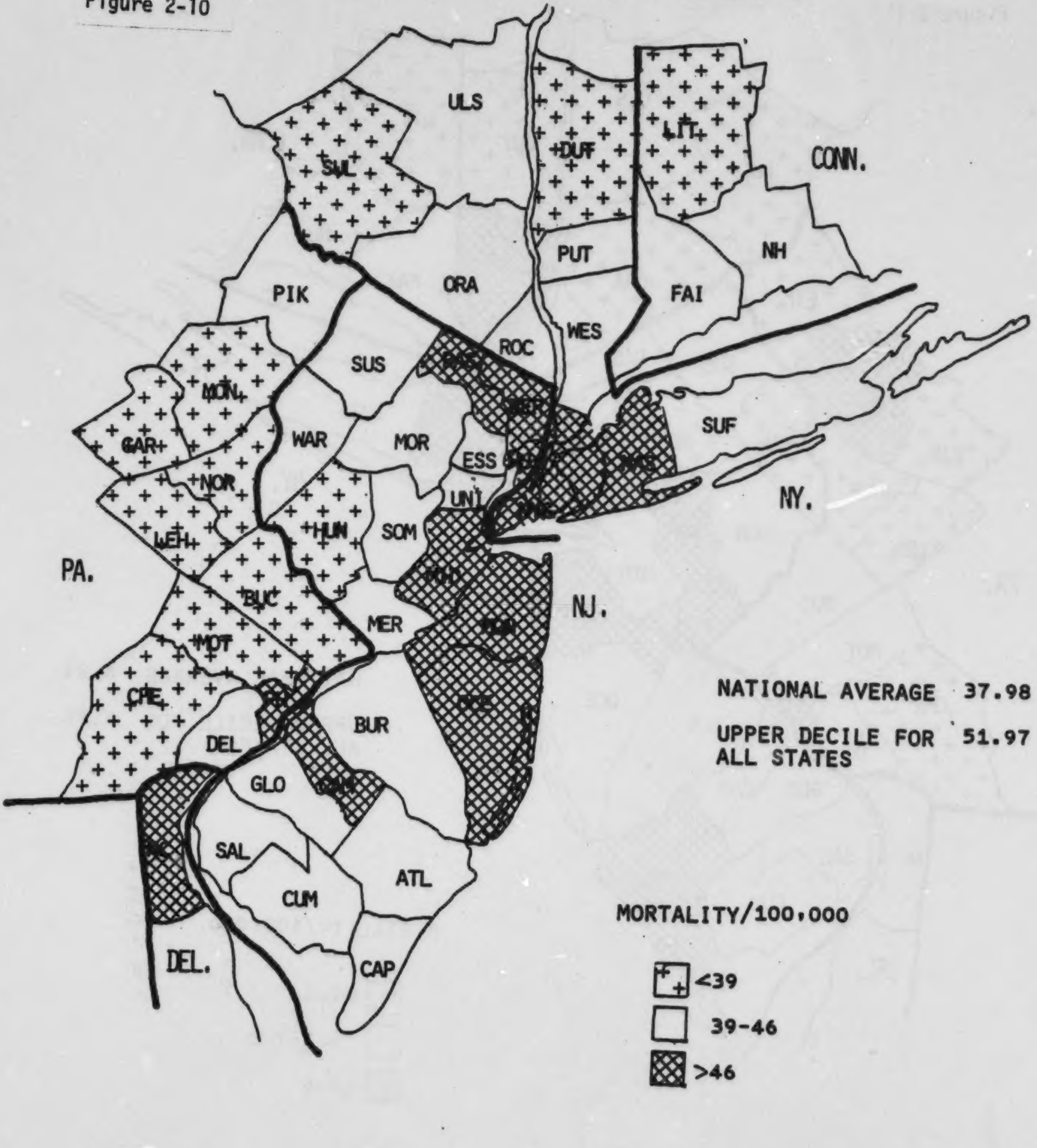
Summarizing, the fifteen maps reproduced in this paper leave the impression that during 1950-1969 the urban-industrial core of the study area had higher rates of cancer mortality than the more rural areas. That picture is in sharpest focus for the male white population and is haziest for the female nonwhite population. A number of the mortality variables seem to covary in spaces. However, judging spatial covariance from maps is subjective. Accordingly, we turned to bivariate correlation analysis.

Correlation Among Cancer Body Site Mortalities

Bivariate correlations were run among the twelve body site and all sites categories for the universe of forty-five political units in the study area. A matrix of 1128 different correlations resulted [48(47)/2]. The degree of geographical coincidence may be judged from a number of perspectives. One approach is to determine a correlation value that will be significantly different from 0 (no correlation) within a specific confidence interval. The value for our study area at 95 percent confidence is .248.

Figure 2-10

MALE WHITE LUNG CANCER MORTALITY, 1950-69

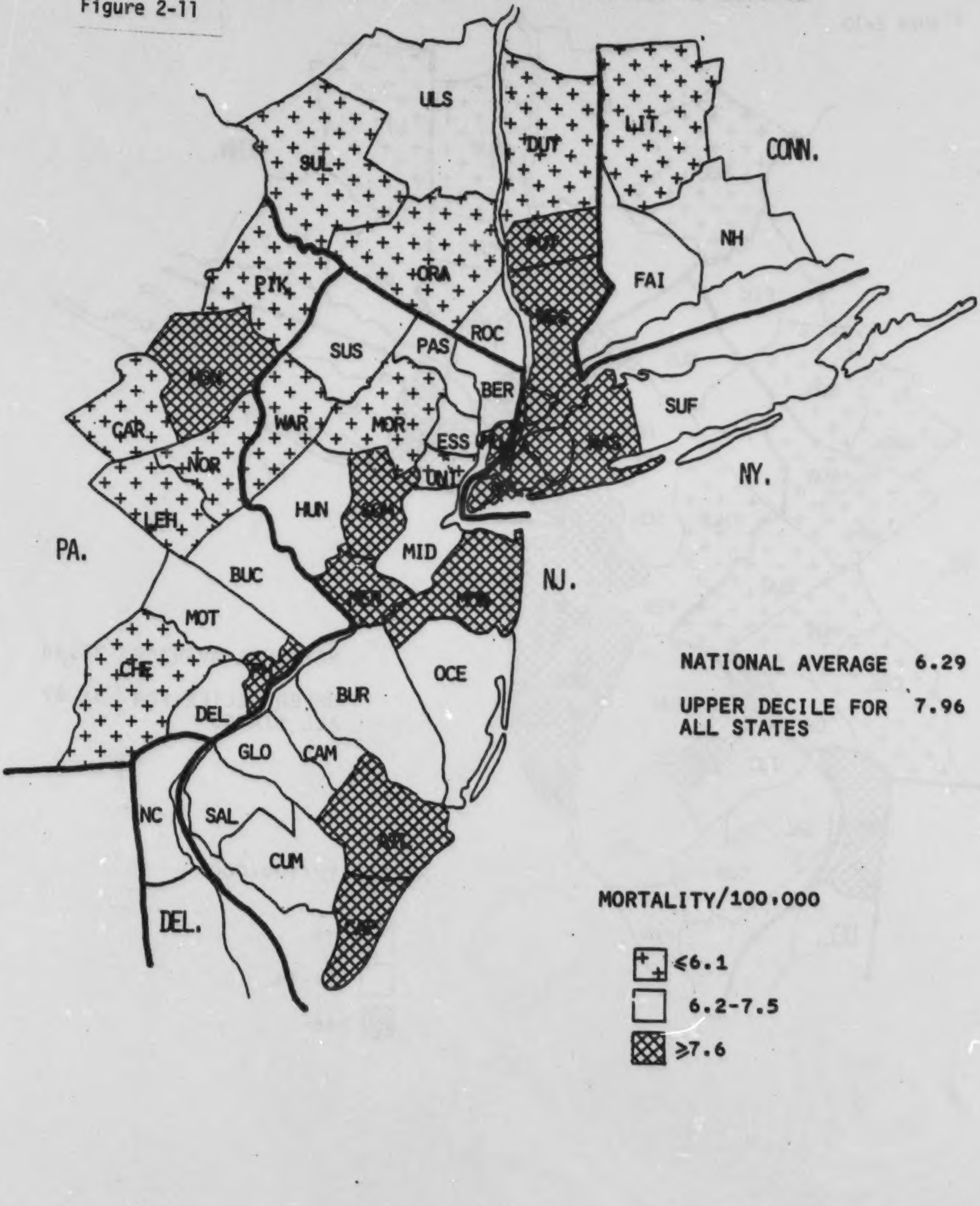

ULS
SUL
DUT
LIT
CONN.
PUT
NH
PIK
ORA
FAI
ROC
WES
SUS
MON
CAR
WAR
MOR
ESS
NOR
SUF
UNI
NY.
HUN
SOM
LEH
PA.
BUC
MER
NJ.
MOT
CHE
BUR
DEL
GLO
SAL
CUM
ATL
DEL.
CAP
MORTALITY/100,000
<39
39-46
>46

NATIONAL AVERAGE 37.98

UPPER DECILE FOR ALL STATES 51.97

Figure 2-11

FEMALE WHITE LUNG CANCER MORTALITY, 1950-69


ULS
SUL
DUT
LIT
CONN.
NH
FAI
ORA
PIK
ROC
PAS
SUS
BER
SUF
CAR
WAR
MOR
ESS
NOR
UNI
HUN
LEH
NY.
PA.
MID
BUC
NJ.
MOT
CHE
OCE
BUR
DEL
GLO
CAM
NC
SAL
CUM
DEL.
MORTALITY/100,000
≤6.1
6.2-7.5
≥7.6

NATIONAL AVERAGE 6.29

UPPER DECILE FOR ALL STATES 7.96

Figure 2-12

FEMALE WHITE BREAST CANCER MORTALITY, 1950-69

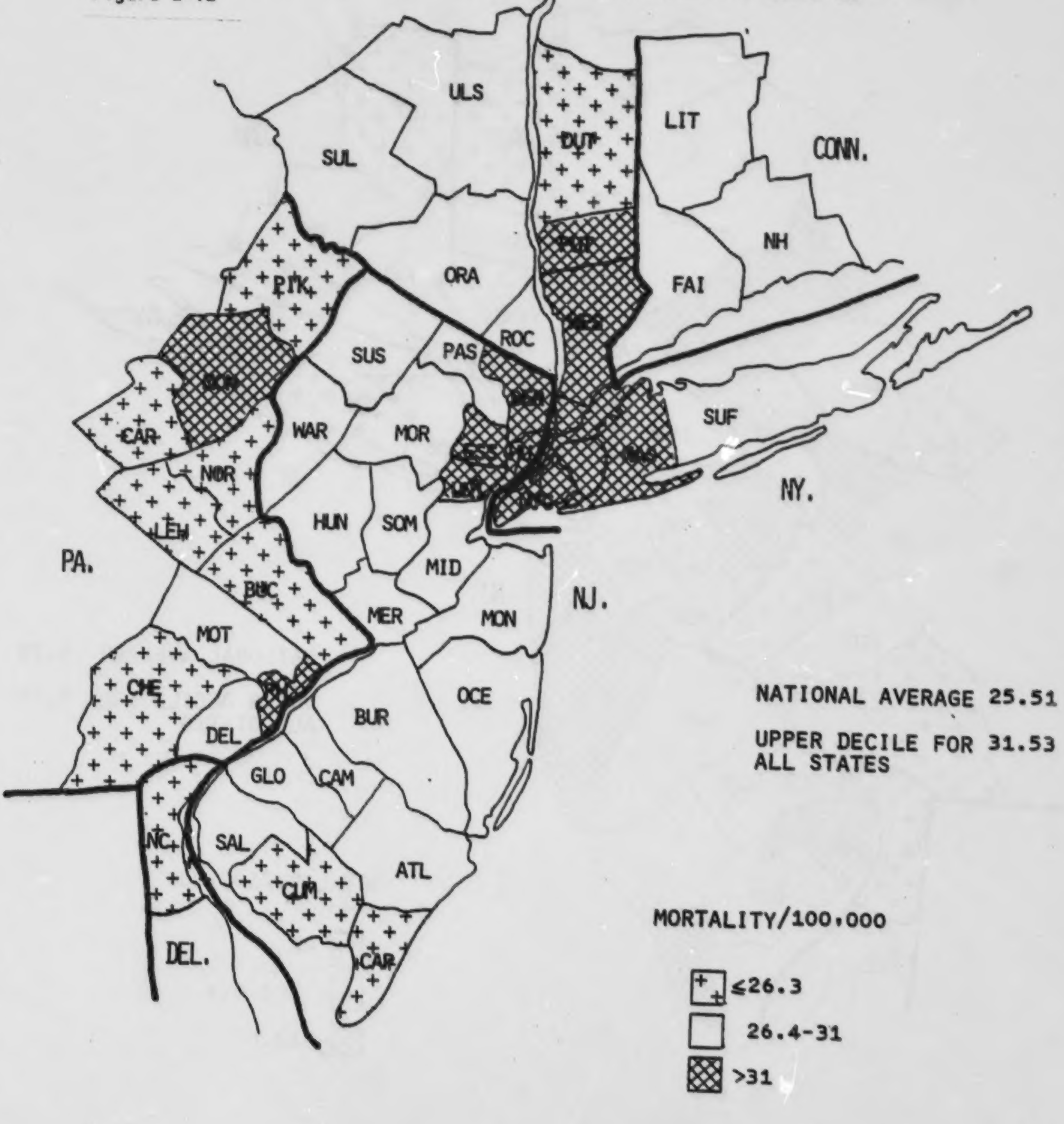

ULS
SUL
LIT
CONN.
DUT
NH
FAI
ORA
PIK
ROC
PAS
SUS
SUF
CAR
WAR
MOR
NOR
NY.
LEH
HUN
SOM
PA.
MID
BUC
NJ.
MER
MON
MOT
CHE
OCE
BUR
DEL
GLO
CAM
NC
SAL
ATL
CUM
CAR
DEL.
NATIONAL AVERAGE 25.51
UPPER DECILE FOR ALL STATES 31.53
MORTALITY/100,000
≤26.3
26.4-31
>31

Figure 2-13

MALE WHITE BLADDER CANCER MORTALITY, 1950-69

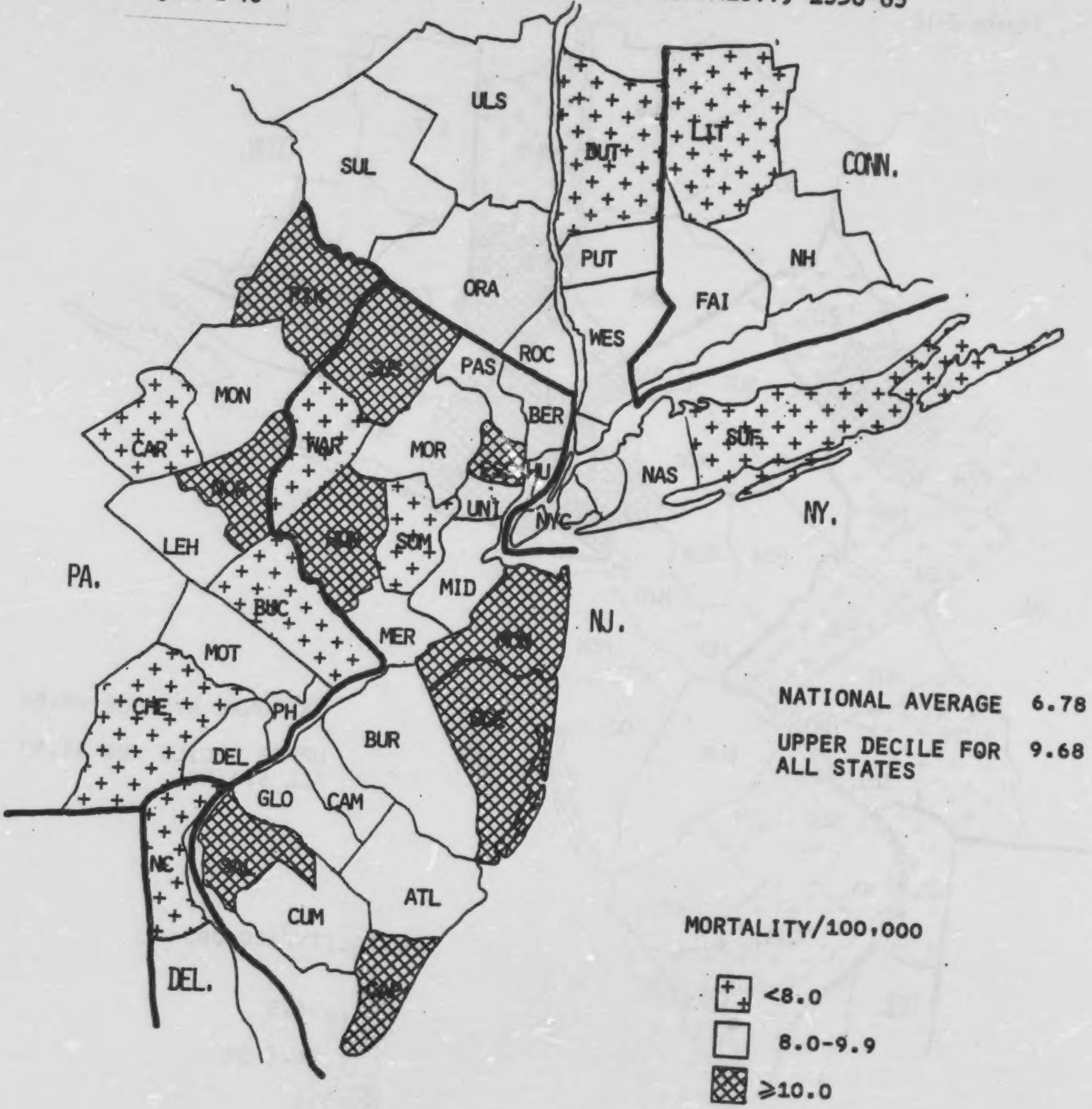

ULS
SUL
DUT
LIT
CONN.
PUT
NH
ORA
FAI
PIK
SUS
WES
ROC
PAS
MON
BER
CAR
WAR
MOR
ESS
HU
NAS
SUF
NOR
UNI
NYC
NY.
HUN
SOM
LEH
PA.
BUC
MID
MER
MON
NJ.
MOT
OCE
CHE
PH
BUR
DEL
GLO
CAM
NC
SAL
ATL
CUM
CAP
DEL.
NATIONAL AVERAGE 6.78
UPPER DECILE FOR ALL STATES 9.68
MORTALITY/100,000
<8.0
8.0-9.9
≥10.0

Figure 2-14

FEMALE WHITE BLADDER CANCER MORTALITY, 1950-69

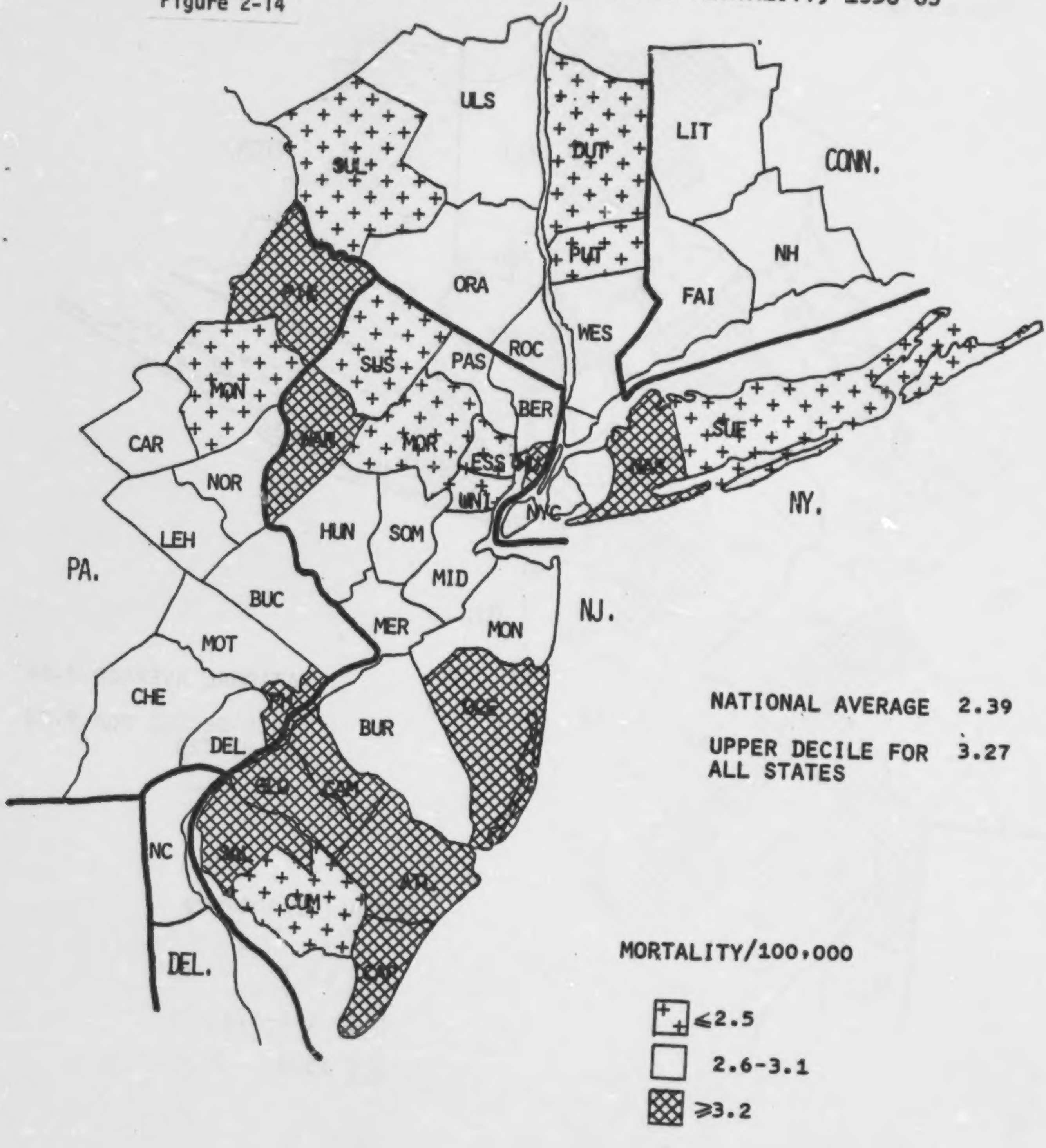

ULS
SUL
DUT
LIT
CONN.
NH
PUT
ORA
FAI
WES
ROC
PAS
SUS
MON
BER
SUF
CAR
MOR
ESS
NOR
UNI
NYC
NY.
LEH
HUN
SOM
PA.
MID
BUC
NJ.
MER
MON
MOT
CHE
BUR
DEL
CAM
NC
CUM
DEL.
NATIONAL AVERAGE 2.39
UPPER DECILE FOR ALL STATES 3.27
MORTALITY/100,000
≤2.5
2.6-3.1
≥3.2

Figure 2-15

MALE WHITE SKIN CANCER MORTALITY, 1950-69

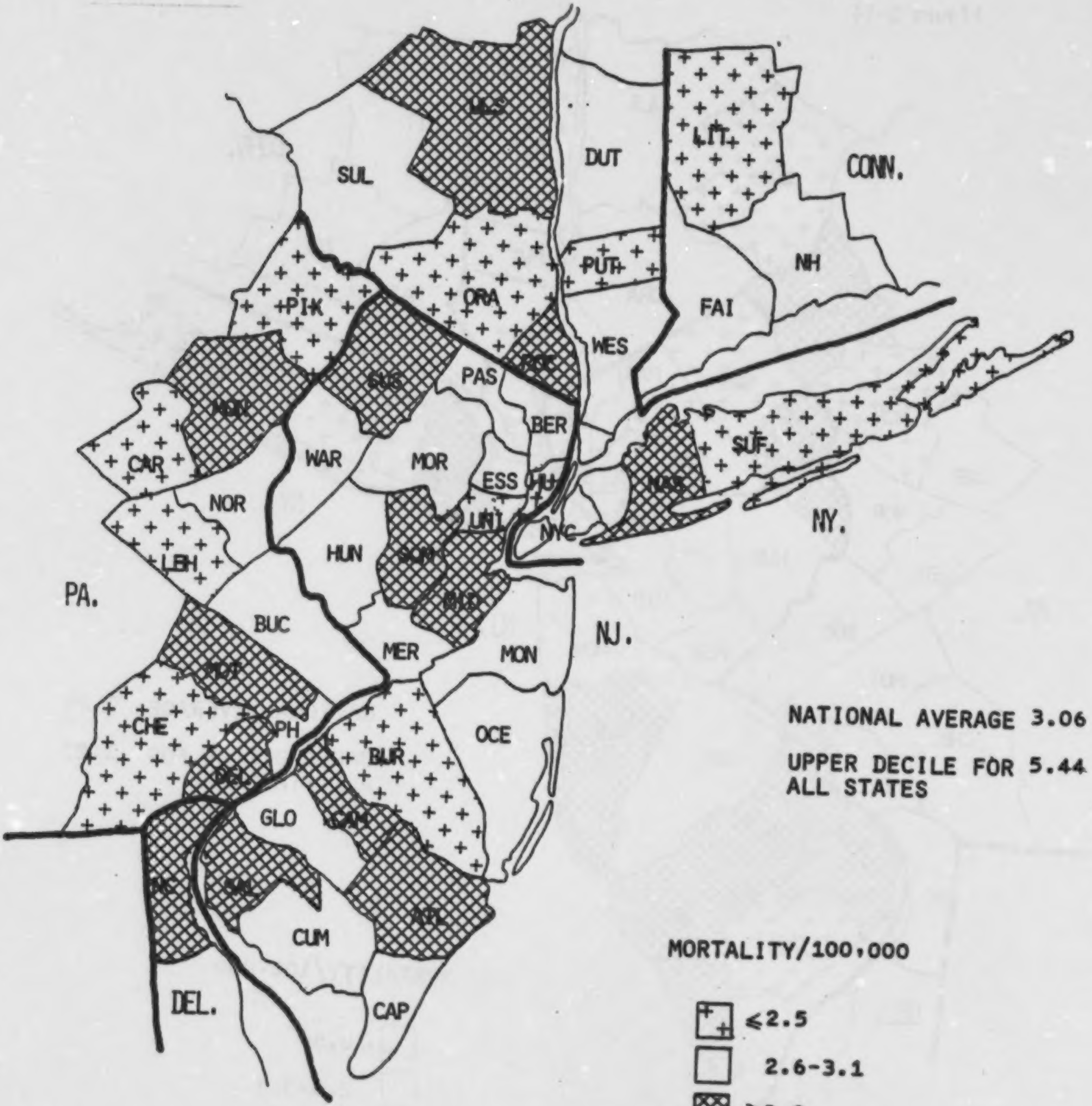
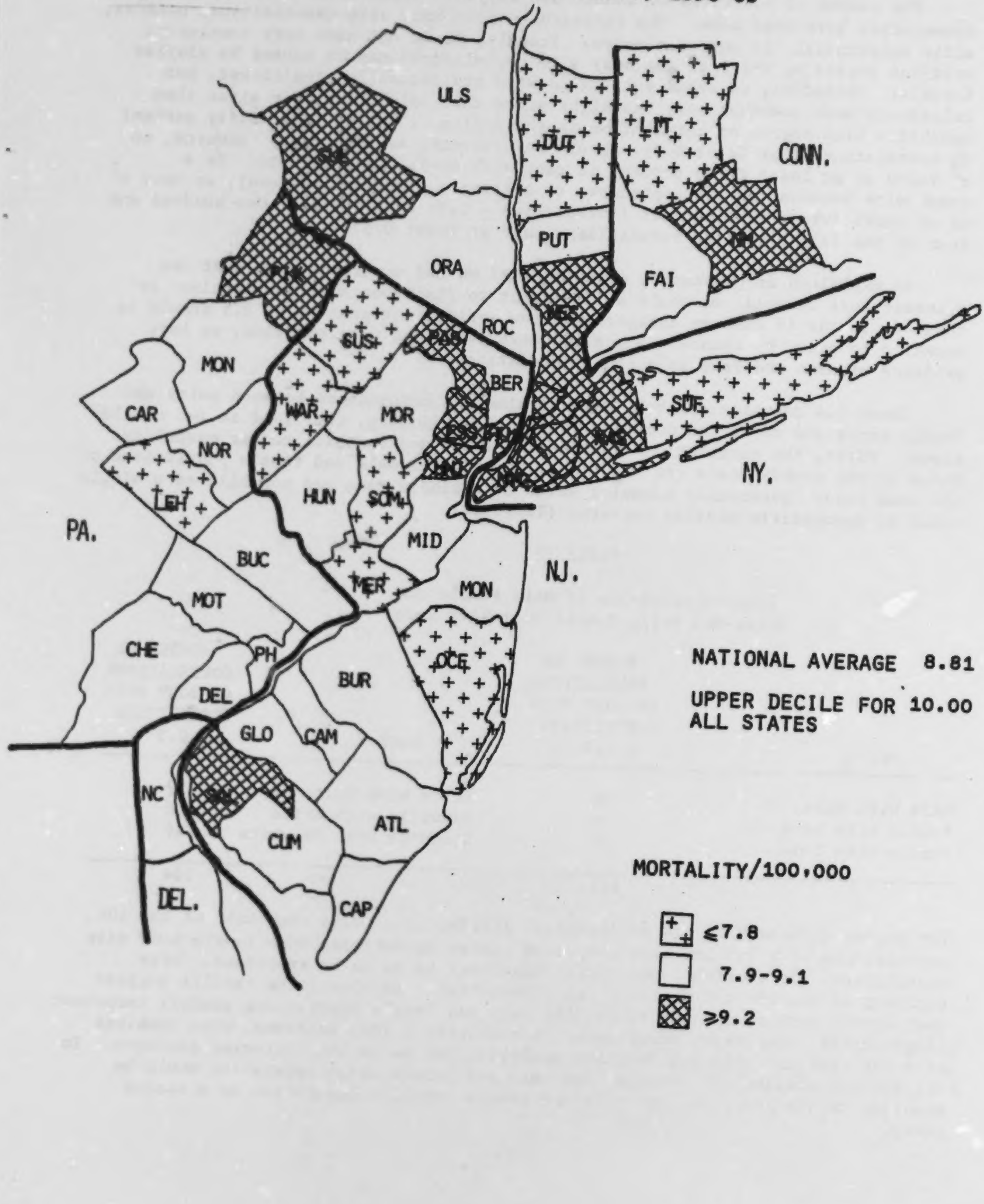

ULS
SUL
DUT
LIT
CONN.
PIK
ORA
PUT
NH
FAI
WES
ROC
SUS
PAS
MON
CAR
BER
SUF
WAR
MOR
ESS
HU
NAS
NOR
UNI
NYC
NY.
LBH
HUN
SOM
MID
PA.
BUC
NJ.
MOT
MER
MON
CHE
PH
BUR
OCE
DEL
GLO
CAM
NC
SAL
ATL
CUM
DEL.
CAP
NATIONAL AVERAGE 3.06
UPPER DECILE FOR ALL STATES 5.44
MORTALITY/100,000
⩽2.5
2.6-3.1
⩾3.2

Figure 2-16 MALE WHITE LEUKEMIA MORTALITY, 1950-69


ULS
DUT
LIT
CONN.
SUL
PIK
PUT
ORA
FAI
NH
WES
ROC
PAS
SUS
MON
BER
SUF
CAR
WAR
MOR
ESS
NAS
NOR
UNI
NY.
LEH
HUN
SOM
PA.
MID
BUC
MER
NJ.
MON
MOT
OCE
CHE
PH
BUR
DEL
GLO
CAM
NC
SAL
ATL
CUM
DEL.
CAP
NATIONAL AVERAGE 8.81
UPPER DECILE FOR 10.00
ALL STATES
MORTALITY/100,000
≤7.8
7.9-9.1
≥9.2

The causes of and cures of cancer are subjects about which many claims of association have been made. The literature about body site associations, however, while substantial, is not conclusive. Frankly, we do not have very conclusive evidence regarding which of the body site correlations may be caused by similar factors. Therefore, to avoid putting forward statistically significant, but relatively weak associations, one might argue that only those body sites that exhibit a high degree of correlation in the region, r^2 of at least fifty percent (a correlation of at least 0.7), should be presented as important. However, an r^2 value of at least fifty percent is rare with social science data. As a compromise between the two approaches to setting a significance level, we used r^2 of at least twenty-five percent (correlation $\geq$ 0.5) as a cutoff. One-hundred and four of the 1128 bivariate correlations were at least 0.5.

As explained in Chapter 1, assuming that we had no expectation that the diseases were related, we could still expect to find some high correlations by chance. In the 48 disease categories above, only 1 correlation of 0.5 should be expected to occur by chance. Since 104 were found and were positive, we have evidence of some moderate to strong associations.

These 104 correlations help us determine the coincidence between males and female rates and white and nonwhite rates. Specifically, they lead to two conclusions. First, the correlations between the white and nonwhite cancer mortality rates at the county scale are not high. Second, the male and female populations of the same races (generously assuming white as a single race and nonwhite as a single race) do demonstrate similar patterns (Table 12).

TABLE 12

Interrelationship of Male-Female and
White-Non White Cancer Mortality Rates

SEX	NUMBER OF CORRELATIONS OF BODY SITE MORTALITIES $\geq$ 0.5	RACE	NUMBER OF CORRELATIONS OF BODY SITE MORTALITIES $\geq$ 0.5
Male with Male	28	White with White	37
Female with Male	58	Nonwhite with White	18
Female with Female	18	Nonwhite with Nonwhite	49
	104		104

The sexual dichotomy is not an important distinction. More than half of the 104 correlations of $\geq$ 0.5 are male body site cancer mortalities with female body site mortalities. In contrast, the racial dichotomy seems to be important. Only eighteen of the 104 correlations are interracial. In short, the results suggest that within each of the two races, the male and female populations exhibit important geographical coincidence among cancer mortalities. This evidence, when combined with the maps and with multivariate analyses, led us to the following decision. In all further statistical studies, the male and female white population would be examined as one group and the male and female nonwhite population as a second group.

With this dichotomy in mind, we concentrated on the intraracial correlations.

The White Population

Over twelve percent of the white population's 276 correlations were at least 0.5; not even one could have been expected to have been at least 0.5 by chance. These thirty-five are listed in Table 13. Two are greater than 0.8 (r^2 is 64 percent), six greater than 0.7 (r^2 = 49%), and fourteen greater than 0.6 (r^2 = 36%). The correlations listed in Table 13 are summarized by body site in Table 14. Table 13 and 14 suggest that many of the internal body cancers covary at the county scale; others do not; and skin and leukemia cancers are not strongly correlated with other body site mortalities. Specifically, mortality from stomach and rectum cancer are strongly correlated with each other and with all cancers. These correlations are particularly interesting because they only comprise about twenty percent of cancer mortalities. In addition, stomach cancer has drastically declined as a cause of cancer mortality.

TABLE 13

Bivariate Correlations ≥ 0.5 Among Thirteen White Population Cancer Mortalities Over 49 County Study Area

$r > .8$	
1 wm* stomach with wf** stomach	.836
2 wm lung with wm all	.802
$r \geq .7$ to $< .8$	
3 wm stomach with wm all	.750
4 wf stomach with wm all	.708
5 wm large intestine with wm all	.789
6 wm all with wf all	.753
$r > .6$ to $< .7$	
7 wm stomach with wm liver	.629
8 wf stomach with wm rectum	.645
9 wf stomach with wf all	.639
10 wm rectum with wf rectum	.638
11 wm rectum with wm all	.640
12 wm rectum with wf all	.621
13 wm liver with wm all	.691
14 wf breast with wf all	.634
$r \geq .5$ to $< .6$	
15 wm stomach with wm rectum	.522
16 wm stomach with wf rectum	.513
17 wm stomach with wf all	.545
18 wf stomach with wm large intestine	.505

TABLE 13 (continued)

$r \geq .5$ to $< .6$ (continued)

19 wf stomach with wm liver	.551
20 wf stomach with wm kidney	.524
21 wm large intestine with wm liver	.592
22 wm large intestine with wm lung	.597
23 wm large intestine with wf breast	.510
24 wm large intestine with wf all	.548
25 wf large intestine with wf all	.529
26 wf rectum with wm liver	.569
27 wf rectum with wm all	.552
28 wf rectum with wf all	.567
29 wm liver with wf all	.504
30 wm lung with wf lung	.571
31 wm lung with wf all	.523
32 wf lung with wf breast	.520
33 wf lung with wf all	.511
34 wf breast with wm all	.592
35 wm kidney with wf all	.558

---

*wm = white male
**wf = white female

TABLE 14

Bivariate Correlations $\geq$ 0.5 Among Thirteen White Population Cancer Mortalities Summarized by Body Site

Body Site	Number of Correlations $\geq$ 0.5
1 Stomach	13
2 Large Intestine	7
3 Rectum	10
4 Liver	6
5 Pancreas	0
6 Lung	7
7 Female Breast	4
8 Male Prostate	0
9 Kidney	2
10 Bladder	0
11 Skin	0
12 Leukemia	0
13 All	21

Lung and large intestine cancer, which account for more than forty percent of male cancer mortalities, also correlate strongly with all cancers and with many of the other body sites. Finally, liver and female breast cancer caused mortalities also covary with a number of other body site mortalities. The lack of strong correlations demonstrated by pancreas, prostate, bladder, skin, and leukemia cancer mortalities implies that in our study area these cancer mortalities may be linked to unique local environments, social conditions, or other factors. Or as Part II of this study will suggest, with the exception of protstate, these weakly correlated diseases exhibit low mortality rates and a high degree of intraregional variation due perhaps to the fact that so few cases have occurred in the study area.

The Nonwhite Population

The 276 nonwhite body site correlations were confusing. On the one hand, about eighteen percent (49) were ≥ 0.5. On the other hand, negative correlations between body sites were found. Upon examination of the raw data, we found that extreme values in ten counties were responsible for the unpredictability of the patterns. For example, Pike County had a nonwhite male all cancer rate of 0 based on no mortalities. Does this mean that nonwhites should all move to Pike County to avoid dying from cancer? Clearly not! The rate for nonwhite females is 299.6/100,000 based on only two recorded mortalities. Similarly, Carbon County's male rate is 537.8/100,000 based on two cases; its female rate is 0 based on no mortalities. To avoid biasing the results by a few cases, ten counties were eliminated from subsequent statistical analyses. Their small contribution to nonwhite cancer mortality in the study area is summarized in Table 15.

TABLE 15

Nonwhite Cancer Mortalities in Ten Counties Eliminated from the Statistical Analysis

County	Number of Nonwhite Cancer Mortalities, 1950-69
1. Litchfield, Conn.	25
2. Putnam, N.Y.	3
3. Sullivan, N.Y.	33
4. Carbon, PA.	2
5. Lehigh, PA.	28
6. Monroe, PA.	15
7. Pike, PA.	2
8. Hunterdon, N.J.	36
9. Sussex, N.J.	30
10. Warren, N.J.	25
Total of 10 Counties	199
Total of Study Area	72,650
Ten Counties % of Study Area	0.27

Removing the ten counties and recalculating the bivariate correlation matrix produced sixteen correlations (5.8 percent of 276) above 0.5. These are listed in Table 16. Table 17 summarizes these by body site.

TABLE 16

Bivariate Correlations ≥ 0.5 Among Thirteen Nonwhite Population Cancer Mortalities Over 35 Political Unit Study Area

*nwm=nonwhite male
**nwf=nonwhite female

≥ .7		
1	nwm* liver with nwf** all	.796
2	nwm lung with nwm all	.779
3	nwf breast with nwf all	.704
≥ .6 to < .7		
4	nwf large intestine with nwf pancreas	.632
5	nwf liver with nwf pancreas	.660
6	nwm lung with nwf all	.656
7	nwm all with nwf all	.673
≥ .5 to < .6		
8	nwf large intestine with nwf liver	.564
9	nwf large intestine with nwf all	.580
10	nwm rectum with nwm liver	.501
11	nwf liver with nwm bladder	.546
12	nwf pancreas with nwf all	.589
13	nwm lung with nwf lung	.519
14	nwm lung with nwm prostate	.505
15	nwm prostate with nwf all	.520
16	nwf kidney with nwf bladder	.517

TABLE 17

Bivariate Correlations $\geq$ 0.5 Among Thirteen Nonwhite Population Cancer Mortalities Summarized by Body Site

	Body Site	Number of Correlations $\geq$ 0.5
1	Stomach	0
2	Large Intestine	3
3	Rectum	1
4	Liver	5
5	Pancreas	3
6	Lung	5
7	Female Breast	1
8	Male Prostate	2
9	Kidney	1
10	Bladder	2
11	Skin	0
12	Leukemia	0
13	All	9

The white and nonwhite body site summaries are quite different. Stomach and rectum cancer mortalities were highly intercorrelated with other body sites among the white population. They are not for the nonwhite population. Whereas pancreas, male prostate and bladder cancer mortalities had no high correlations among the white population, they do have some important correlations among the nonwhite populations.

A useful follow-up to testing the correlation of pairs of body sites at the county level would be to search the literature for evidence of casual effects linking the body sites.

Studying pairs of body site mortalities is a relatively simple means of studying coincidence. If some pairs of cancer mortalities are related, we would suspect that groups of 3, 4, 5, etc. body sites might also be related. We, therefore, next turn to multivariate statistical analyses.

Patterns of White Population Cancer Mortality Revealed by Factor Analyses of Diseases

Factor analysis is a statistical tool for measuring the extent to which groups of different indicators intercorrelate. With respect to cancer mortality data, factor analysis will classify groups of counties according to groups of cancer mortalities that they strongly manifest. It does this by creating statistical cancer indicators that are identified with particular counties. Factor analysis has many options that imply that the user must be extremely careful in applying the technique. The factor analysis results that we present have the following characteristics: (1) they reappeared throughout the testing procedure: (2) they demonstrated a distinctive geographical pattern.

Two types of factor analytic models are presented in this study: unrotated and varimax rotated. The unrotated solution is the simplest geometry in which the first factor accounts for the most common variation, the second factor the second most, and so on. The advantage of the unrotated solution is that few factors are produced and therefore have to be interpreted. The disadvantage is that the few factors are sometimes too general to be useful. By rotating the unrotated solution, the analyst generally produces more factors and factors that are more distinctive than the unrotated factors. The varimax rotation is one of a set of rotations that may be used. In the author's experience, it is the most widely used because the factors it produces tend to be useful.

The factor analyses of the white cancer mortality data distinguished five groups (Figure 17) of counties with the following characteristics:

1. Located in the urban-industrial core; exhibit high rates of all, stomach, male large intestine, rectum, lung, male liver, and female breast cancer mortality;

2. Located in relatively rural areas; demonstrate low rates of all types of cancer mortality;

3. Centered in southern New Jersey, but with other outliers; exhibit high rates of bladder, male leukemia and female large intestine cancer mortalities;

4. Centered in northeastern Pennsylvania and in adjacent areas in New Jersey and New York; high rates of female mortalities from kidney, skin, lung and leukemia;

5. Located throughout the region; do not clearly identify with any of the above four groups.

Urban Industrial Cancer Corridor and Low Cancer Mortality Groups

These two groups were primarily distinguished by a single factor that on a multivariate basis defines the high cancer rate urban-industrial corridor zone and contrasting low cancer mortality counties. This factor accounted for about thirty percent of the covariation within the matrix of white cancer mortality information. The important factor loadings of an unrotated principal components solution and a varimax rotated solution are listed below (Table 18). The factor loadings are correlations of the variable (all white male, all female white, etc.) with the statistically created urban-industrial cancer corridor variable.

The unrotated solution pivots around high rates of all cancers and stomach cancer mortalities. Five other body sites also make important contributions to defining the factor. The factor is split among male and female loadings. The rotated version focuses on male lung, all sites and stomach cancer. All eleven mortality types contribute significantly to both factors. In essence, this factor is a statistical aggregation of the maps and bivariate correlations that indicated strong correlations among many of the body sites focused in the urban-industrial corridor.

WHITE POPULATION COMPOSITE CANCER MORTALITY GROUPS, 1950-69

Figure 2-17

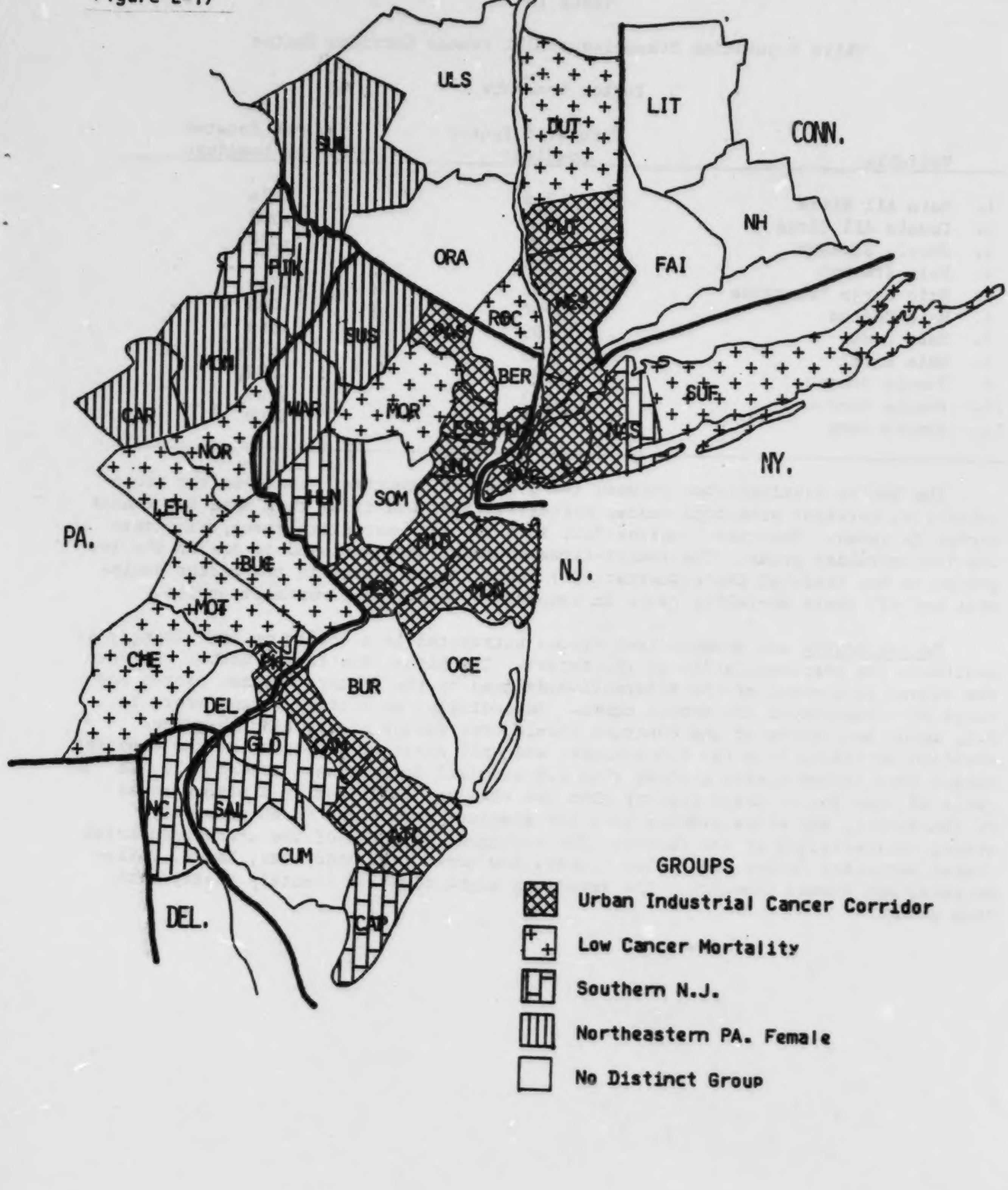

ULS
DUT
LIT
CONN.
SUL
PUT
NH
ORA
FAI
PIK
ROC
WES
SUS
MON
BER
CAR
MAR
MOR
SUF
NOR
ESS
NAS
NY.
UNI
LEH
HUN
SOM
PA.
MID
BUC
NJ.
MER
MON
MOT
CHE
OCE
BUR
DEL
GLO
CAM
NC
SAL
ATL
CUM
DEL.
CAP
GROUPS
Urban Industrial Cancer Corridor
Low Cancer Mortality
Southern N.J.
Northeastern PA. Female
No Distinct Group

TABLE 18

White Population Urban-Industrial Cancer Corridor Factor

Factor Loadings

Variable	Unrotated Factor Loadings	Varimax Rotated Factor Loadings
1. Male All Sites	.946	.824
2. Female All Sites	.894	.585
3. Female Stomach	.789	.520
4. Male Stomach	.764	.521
5. Male Large Intestine	.722	.789
6. Male Rectum	.700	.486
7. Male Lung	.689	.845
8. Male Liver	.683	.658
9. Female Breast	.659	.584
10. Female Rectum	.653	.606
11. Female Lung	.493	.709

The factor distinguishes between two groups of counties: those in the urban-industrial corridor with high cancer mortality rates and those that have low cancer mortality rates. Fourteen counties fall into the high mortality group, nine into the low mortality group. The twenty-three counties were included in one of the two groups on the basis of two criteria: (1) their factor scores in the factor analysis; and (2) their mortality rates in comparison to national mortality rates.

Factor scores are standardized values attributed to a county on how strongly it manifests the characteristics of the factor. The higher the factor score, the more the county is a model of the information defined by the factor. Factor scores are based on values under the normal curve. Accordingly, an average factor score is 0.0, about two-thirds of the counties should have factor scores less than one standard deviation from the 0.0 average, and only about five percent of the counties should have factor scores greater than 2.0 standard deviations. Accordingly, in Table 19, any factor score greater than one standard deviation is a strong model of the factor, any score greater than two standard deviations is an extremely strong manifestation of the factor. The strongest examples of the urban-industrial cancer mortality factor are Hudson County, New York City, Middlesex, Essex, Philadelphia, and Nassau Counties. The remaining eight counties identify weakly with this group.

TABLE 19

Counties Belonging to the Urban-Industrial Cancer Corridor Group

(in order of their ranking)

		FACTOR SCORE		NUMBER OF CANCER MORTALITY RATES ≥ UPPER DECILE FOR STATES	
	COUNTY	UNROTATED FACTOR	VARIMAX ROTATED FACTOR	11 TYPES DEFINING FACTOR	ALL 24 MORTALITY VARIABLES
1	Hudson	2.20	2.37	9	10
2	New York City	2.03	1.32	10	11
3	Middlesex	1.56	1.37	7	8
4	Essex	1.50	1.22	7	9
5	Philadelphia	1.67	1.76	5	6
6	Nassau	1.40	1.11	4	8
7	Putnam	0.69	1.50	5	7
8	Passaic	1.15	0.58	5	6
9	Westchester	0.63	0.76	5	5
10	Camden	0.60	0.78	4	7
11	Mercer	0.61	0.61	4	5
12	Monmouth	0.37	0.83	4	7
13	Union	1.01	0.09	4	5
14	Atlantic	0.36	0.69	5	6

The focus of the urban-industrial cancer corridor factor on the eleven specific mortalities is closely supported by the comparison between each county's rates and the upper decile rate of the states. For example, New York City exceeds the upper decile for the nation in eleven of the twenty-four mortality categories. Ten of the eleven categories in which New York City exceeds the upper decile for the nation are defined by the urban-industrial cancer corridor factor. For Hudson County, nine of the ten upper decile rates that it exceeds are included among the eleven variables in this factor. As the reader looks down the page the correspondence between the factor scores and the county to national upper decile rate continues. Mercer, Monmouth, Union, and Atlantic Counties clearly identify with the indicators of the urban-industrial cancer corridor group. However, their identification is not nearly as strong as the other ten counties.

In contrast to the high cancer mortality counties, the low cancer mortality counties are located on the region's periphery. Nine counties form the essence of this factor. They are characterized by high negative identification with the cancer mortalities characteristic of the urban-industrial cancer corridor counties. Furthermore, rarely do these counties exceed the upper decile for the states in any of the twenty-four cancer mortalities included in this study. On the contrary, they frequently exhibit mortality rates below the national average (Table 20).

TABLE 20

Counties Belonging to the Low Cancer Mortality Group (in order of their ranking)

		FACTOR SCORE		NUMBER OF CANCER MORTALITY RATES	
	COUNTY	UNROTATED FACTOR	VARIMAX ROTATED FACTOR	≥ UPPER DECILE FOR STATES FOR 24 TYPES	≤ NATIONAL AVERAGE FOR 24 TYPES
1	Chester	2.03	1.59	0	16
2	Dutchess	2.14	1.10	0	17
3	Bucks	1.36	1.13	0	10
4	Montgomery	1.14	1.00	0	8
5	Lehigh	0.97	1.02	1	9
6	Northampton	0.68	0.92	1	7
7	Suffolk	1.47	0.07	0	12
8	Rockland	1.06	0.34	1	12
9	Morris	0.77	0.24	2	7
10	Orange	0.18	0.04	2	7
11	Ulster	0.03	0.01	2	4

In addition to these nine counties, Orange and Ulster Counties also identified weakly with the low cancer mortality group. These quantitative values are also summarized in the Table 20. Ten of the eleven counties in this group share a common characteristic. They lie west of the major regional thoroughfares and sources of air pollution. The exception is Suffolk County. None of Suffolk's cancer mortality rates exceed the upper decile for the states and half of its rates are below the national average. Suffolk, however, does not identify strongly with the rotated factor analysis. It is therefore not included as one of the five exemplary low mortality group counties. It does not identify with the varimax rotated factor because two of the four most important body sites defining the rotated factor are lung cancer. Suffolk, which lies in the path of the prevailing southwesterlies that carry air pollution from the urban-industrial corridor to the northeast, has lung cancer mortality rates of 44.6 and 7.3 for the male and female white populations, respectively. The difference between lung cancer mortality rates in Suffolk and the other eight counties that model this factor is marked. The average lung cancer rates for these counties are 36.6 and 6.0, respectively. Suffolk County's lung cancer mortality rate exceeds those of the other eight low cancer mortality counties by more than twenty percent.

Perhaps the best way of summarizing the urban industrial cancer corridor factor is to compare the counties which most strongly manifest its statistical composition (Table 21). The six urban-industrial corridor counties exhibit male cancer mortality rates twenty-six percent higher than the national average and female rates nineteen percent higher than the national average. Cancer mortality rates in the five low mortality counties are slightly below national average levels.

TABLE 21

Comparison of Cancer Mortality Rates in Counties with Extremely Different Rates

		Mortality Rate/ 100,000		County Average / National Average Mortality Rate/ 100,000		Mortality Rates
	County	White Male	White Female	White Male	White Female	≥ Upper Decile for the States: 24 types
1	Hudson	232	154	1.33	1.18	10
2	New York City	216	160	1.24	1.23	11
3	Middlesex	221	149	1.27	1.15	8
4	Essex	215	155	1.24	1.19	9
5	Philadelphia	221	156	1.27	1.20	6
6	Nassau	212	155	1.22	1.19	8
7	Average of 1-6	220	155	1.26	1.19	8.7
8	Chester	158	131	.91	1.01	0
9	Dutchess	167	119	.96	.91	0
10	Bucks	174	131	1.00	1.01	0
11	Northampton	188	134	1.08	1.03	0
12	Lehigh	178	131	1.02	1.01	1
13	Average of 8-12	173	129	.99	.99	0.2
14	Row 7 / Row 13	1.27	1.20	1.27	1.20	—

Southern New Jersey Bladder, Female Large Intestine, and Male Leukemia Factor

The third group identified by the factor analysis is concentrated in southern New Jersey, but has important outliners elsewhere. Seven of the nine counties are located in southern New Jersey and adjacent Philadelphia and New Castle Counties. The two counties not in the general southern New Jersey area are Pike County, Pennsylvania and Nassau County, New York.

The factor accounts for about ten percent of the information in the cancer mortality bivariate intercorrelation matrix. The four important factor loadings are listed in Table 22. Female bladder cancer mortality is the highest loading in the rotated and unrotated versions; male leukemia identifies weakly in both versions.

TABLE 22

White Population Southern New Jersey Bladder, Female Large Intestine, and Male Leukemia Factor

	Variable	Unrotated Factor Loadings	Varimax Rotated Factor Loadings
1.	Female Bladder	.718	.831
2.	Male Bladder	.579	.626
3.	Female Large Intestine	.527	.688
4.	Male Leukemia	.506	.484

Summary data for the nine counties are presented below (Table 23). Pike, Salem, and Gloucester Counties epitomize the factor. Philadelphia, Nassau, and Cape May Counties strongly identify with the factor. The identification of Camden, Hunterdon, and New Castle counties with the cancer mortalities in the factor is much weaker.

TABLE 23

Counties Belonging to the Southern New Jersey Bladder, Female Large Intestine, and Male Leukemia Factor

(in order of their ranking)

		FACTOR SCORE		WITH RESPECT TO FOUR CANCER MORTALITIES	
	COUNTY	UNROTATED FACTOR	VARIMAX ROTATED FACTOR	≥ TO UPPER DECILE FOR THE STATES	≥NATIONAL AVERAGE FOR THE STATES
1	Pike	4.20	3.63	4	4
2	Salem	1.97	1.81	2	4
3	Gloucester	1.17	1.40	1	4
4	Philadelphia	0.55	1.15	1	3
5	Nassau	1.19	0.50	2	4
6	Cape May	0.81	0.80	2	4
7	Camden	0.22	0.72	2	3
8	New Castle	0.36	0.14	0	3
9	Hunterdon	-0.06	0.96	2	3

Inexplicably, Pike County, Pennsylvania is the strongest manifestation of the factor. Preliminary investigations suggest that high mortalities in this county may be due to the relocation of migrants from the southern New Jersey area. With the exception of Nassau County, New York, the other counties are located in southern

New Jersey and environs. Chapter 5 will review what we learned about this factor and Part II of this study will suggest an ethnic explanation for this factor.

Northeastern Pennsylvania Female Cancer Mortality

The final factor isolated a group of counties in northeastern Pennsylvania and adjacent portions of New York State and New Jersey that exhibit above average rates of female cancer mortality. The factor loadings (Table 24) and factor scores (Table 25) describe a female population with relatively high rates of kidney, skin, and leukemia cancer and low lung cancer rates. In the rotated factor analysis, the factor split into two factors.

Female cancer mortality rates in these seven counties are twelve percent above the national average, while male rates in the same counties are only five percent above the national central tendency. These seven counties do not manifest the highest female mortality rates in the region. However, they do demonstrate high rates of female mortality relative to their male population rates. These mortalities are associated with body sites that differ from those most strongly identified with the urban-industrial cancer corridor counties. Their relatively low lung cancer mortality rates are another feature that distinguishes these seven counties form the counties in the urban-industrial group. This factor accounted for only 7.5 percent of the correlation matrix covariance. The Northeastern Pennsylvania factor could be related to occupational factors, life-style, and it could be statistical creation with no real significance. Accordingly it will require substantial research to clarify. Chapter 5 describes what we learned about this factor, and Part II of this report describes a changed version of this factor for the time period 1968-72.

TABLE 24

Northeastern Pennsylvania White Female Population Cancer Factor

Variable	Unrotated Factor Loadings	Rotated Factor Loadings A	Rotated Factor Loadings B
1. Kidney	.694	.766	—
2. Skin	.485	.757	—
3. Lung	-.480	—	—
4. Leukemia	.442	—	.590
5 Liver	—	—	.775

Other Body Sites and Counties

Five of the twenty-four sex/body site cancer mortalities and eleven counties did not identify strongly with one of the four groups that have been reviewed. The sex/body sites combinations are: male and female pancreas; male prostate; male kidney; and male skin. For the reader who seeks further information about the body site correlations, Appendix A provides four sets of information: (1) the means and standard deviations of the white population data; (2) the bivariate correlations significant at the .05 level; (3) the multiple correlation of each of the 24 cancer types; and (4) the communality (covariance accounted for by factor analysis) of each of the 24 variables.

TABLE 25

Counties Belonging to Northern Pennsylvania Female Cancer Mortality Factor

(in order of their ranking)

	County	Unrotated Factor	Factor Score Varimax Rotated Factors A	B	Ratio of County To National Avg. Female	Male
1.	Sussex	3.13	4.40	0.31	1.08	1.04
2.	Sullivan	2.21	0.67	0.25	1.11	1.08
3.	Carbon	1.84	-0.10	3.95	1.07	1.01
4	Hunterdon	1.06	0.59	1.72	1.10	1.01
5	Pike	1.74	0.43	-1.37	1.15	1.06
6	Warren	0.74	1.68	-0.33	1.14	1.09
7	Monroe	-0.16	-1.36	0.92	1.21	1.05
8	Average 1-7	–	–	–	1.12	1.05

Eleven of the forty-five political units did not closely identify with any of the four groups: Fairfield, Litchfield, New Haven, Bergen, Burlington, Cumberland, Ocean, Somerset, Orange, Ulster, and Delaware.

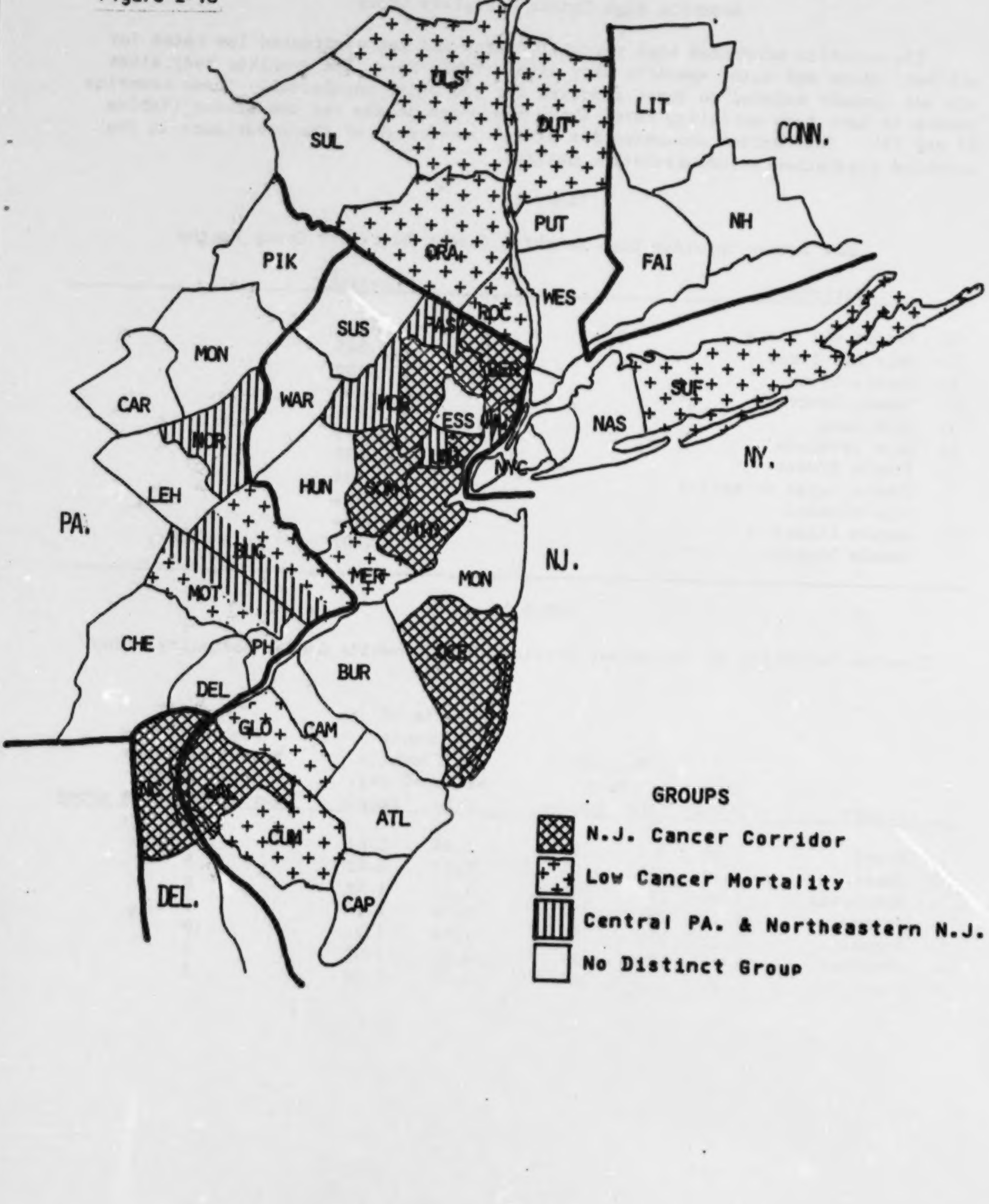
Patterns of Nonwhite Population Cancer Mortality Revealed by Factor Analysis

The nonwhite population data was factor analyzed for a thirty-five county study area. Four groups of counties resulted. The patterns are not as distinct as those uncovered among the white population. Many of the counties were located in more than one group. More than forty percent of the counties (14 of 35) could not be classified. Added to the ten counties that were eliminated from the original forty-five units we have been able to classify less than half of the political units (Figure 18). The following is an overview of the four groups:

1. Located in northeastern and southwestern New Jersey; high rates of many types of cancer mortality;
2. Located in New York and Pennsylvania; demonstrate low rates of all cancer mortalities;
3. Concentrated in central Pennsylvania counties of the region and in northerneastern New Jersey; exhibit high rates of specific male cancer mortalities;
4. Dispersed throughout the region, do not identify with the other four factors.

NONWHITE POPULATION COMPOSITE CANCER MORTALITY GROUPS, 1950-69

Figure 2-18


ULS
DUT
LIT
CONN.
SUL
PUT
NH
ORA
PIK
FAI
ROC
WES
PAS
SUS
MON
BER
SUF
CAR
WAR
MOR
ESS
NAS
NOR
NYC
NY.
LEH
HUN
SOM
PA.
MID
BUC
NJ.
MER
MOT
MON
CHE
PH
OCE
BUR
DEL
GLO
CAM
NC
SAL
ATL
CUM
DEL.
CAP
GROUPS
▩ N.J. Cancer Corridor
+ Low Cancer Mortality
▥ Central PA. & Northeastern N.J.
☐ No Distinct Group

New Jersey Corridor
Nonwhite High Cancer Mortality Group

Ten counties exhibited high mortality rates and ten manifested low rates for all body sites and eight specific body sites (Table 26). The specific body sites are not closely related to those reported for the white population. These counties tended to have many mortality rates above the upper decile for the states (Tables 27 and 28). This factor accounted for almost one-fourth of the covariance in the nonwhite population intercorrelation matrix.

TABLE 26

New Jersey Corridor High Nonwhite Cancer Mortality Group Factor

Variable	Loadings
1. Female All Sites	.914
2. Male All Sites	.845
3. Female Liver	.700
4. Female Pancreas	.686
5. Male Lung	.674
6. Male Prostate	.655
7. Female Breast	.652
8. Female Large Intestine	.584
9. Male Bladder	.474
10. Female Kidney	.464
11. Female Bladder	.407

TABLE 27

Counties Belonging to New Jersey Corridor High Nonwhite Cancer Mortality Group

County	Factor Score	Mortality Rate Male	Mortality Rate Female	Ratio of County To Nonwhite National Avg. Male	Ratio of County To Nonwhite National Avg. Female	Number of Mortality Rates ≥ Upper Decile For States: 24 types
1. Ocean	1.94	265	232	1.44	1.67	11
2. Hudson	1.50	290	197	1.57	1.42	9
3. Middlesex	1.40	279	208	1.51	1.49	8
4. Bergen	1.24	281	192	1.52	1.38	6
5. Passaic	1.22	272	183	1.48	1.31	10
6. Somerset	1.71	227	229	1.23	1.65	6
7. Union	.82	252	166	1.37	1.19	5

TABLE 27 (continued)

Counties Belonging to New Jersey Corridor High Nonwhite Cancer Mortality Group

	County	Factor Score	Mortality Rate Male	Mortality Rate Female	Ratio of County To Nonwhite National Avg. Male	Ratio of County To Nonwhite National Avg. Female	Number of Mortality Rates ≥ Upper Decile For States: 24 types
8.	New Castle	.79	269	167	1.46	1.20	5
9.	Salem	.60	232	162	1.26	1.16	9
10.	Morris	.57	249	170	1.35	1.22	8
11.	Average 1-10		262	191	1.42	1.37	

The ten high mortality counties exhibit rates much higher than the national average. Nine of the ten counties are in New Jersey, and the dominant impression is concentration along the New Jersey Turnpike. New Castle County, the exception to the New Jersey dubious distinction of monopolizing this group, lies immediately adjacent to the southern New Jersey extension of this group. Ocean County does not contain the New Jersey Turnpike. However, the Garden State Parkway runs through Ocean County. It is a major throroughfare that not only adds automobile emissions but also has opened Ocean County to migration from northeastern New Jersey. Although we have been told that Ocean County's black population has long resided in the county, the Ocean County case suggests intrastate migration.

A surprising finding was that three of the most urban-industrial political units (New York City, Philadelphia, and Essex County) did not closely identify with this group. At this time, we hypothesize that occupational differences may account for this difference. Specifically, a larger proportion of nonwhite males in New Jersey may be engaged in occupations exposing them to industrial hazards.

Low Mortality Counties

The low mortality counties are primarily located in New York State and Pennsylvania (Figure 18). Three are, however, in central and southern New Jersey. The strongest manifestations of this group are Suffolk, Dutchess and Rockland Counties in New York State (Table 28). Mercer and Montgomery Counties identify less strongly with this low mortality group.

The difference between the high and low mortality nonwhite groups is more striking than the differences between the high and low mortality white population groups. The ten high nonwhite mortality counties average greater than forty percent more deaths from cancer than the low mortality group. The distinction between these two groups, like their white population counterparts, is almost entirely due to the extremely high rates in the high cancer mortality rate counties.

In essence, counties that exhibit average cancer mortality rates from the national perspective are low mortality counties in this study area. The high rate counties in the study area are close to or above the upper decile rates for many of the body sites.

Central Pennsylvania and Northeastern New Jersey Nonwhite Male Liver, Rectum, and Kidney Cancer Group

A third group of counties, focused in the central Pennsylvania counties of the study area and in northeastern New Jersey, exhibit relatively high rates of five types of cancers and all male neoplasms. The factor loadings (Table 29) point to five body sites that were not involved in the New Jersey corridor cancer factor and repeat the male all sites variable.

TABLE 28
Counties Belonging to Nonwhite Low Mortality Group

	County	Factor Score	Mortality Rate Male	Mortality Rate Female	Ratio of County To Nonwhite National Avg. Male	Ratio of County To Nonwhite National Avg. Female	Number of Mortality Rates ≤ Upper Decile For States: 24 types
1.	Suffolk	-1.92	132	93	.72	.67	20
2.	Dutchess	-1.86	147	195	.80	1.40	19
3.	Rockland	-1.59	148	104	.80	.75	14
4.	Ulster	-1.64	175	118	.95	.85	13
5.	Bucks	-1.22	225	116	1.22	.83	14
6.	Cumberland	-1.02	156	144	.85	1.03	15
7.	Orange	- .86	204	110	1.11	.79	15
8.	Gloucester	- .83	183	146	.99	1.05	13
9.	Mercer	- .32	204	150	1.11	1.08	10
10.	Montgomery	- .20	223	124	1.21	.89	9
	Average 1-10	-	180	130	.98	.93	-

TABLE 29

Central Pennsylvania and Northeastern New Jersey Nonwhite Male Liver, Rectum, and Kidney Cancer Factor

	Variable	Factor Loadings
1.	Male Liver	.673
2.	Male Rectum	.641
3.	Male Kidney	.568
4.	Female Lung	.488
5.	Male All Sites	.459
6.	Male Skin	.458

Four of the seven counties also identified strongly with the New Jersey nonwhite corridor factor (Table 30). Again occupational exposure is suggested as a hypothesis for the geographic concentration of the group.

TABLE 30

Counties Belonging to the Central Pennsylvania and Northeastern New Jersey Nonwhite Male Liver, Rectum and Kidney Cancer Factor

		Of Six Mortality Variables	
County	Factor Score	County ≥ Upper Decile for States	County ≥ National Average
1. Northampton	2.08	4	6
2. Bucks	1.75	2	5
3. Hudson	1.47	3	6
4. Passaic	1.25	2	6
5. Union	.95	2	5
6. Morris	.90	4	6
7. Montgomery	.70	0	4

While Bucks and Montgomery Counties have had low enough female cancer mortalities to place them into the low cancer mortality group, the nonwhite male populations in these counties have died from a number of specific body site neoplasms.

Other Body Sites and Counties

Six of the twenty-four sex/body site cancer mortalities and 14 of the 35 counties did not strongly identify with one of the three groups. The six sex/body site combinations are: male and female stomach, male large intestine, female rectum, female skin and male leukemia. Appendix B 2 provides four sets of additional information about each of 24 cancer types: (1) the means and standard deviations of the thirty-five county nonwhite population data; (2) the bivariate correlations significant at the .05 level; (3) the multiple correlation of each of the 24 cancer types; and (4) the communality of the 24 sex/body site types.

Fourteen of the thirty-five political units did not clearly identify with any of the three groups: Fairfield, New Haven, Atlantic, Burlington, Camden, Cape May, Essex, Monmouth, Nassau, New York City, Westchester, Chester, Delaware, and Philadelphia. Overall, given the relatively few number of nonwhite mortalities in the study area, the fact that nonwhite consists of many races, and the high inter and intraregional migration rates of blacks, we do not have enough confidence in the data to do any further statistical studies of the nonwhite population.

Summary

This chapter has studied the spatial distribution of cancer mortalities in a forty-nine county study area. The next two chapters will present the association between cancer mortality and high and low risk factors.

CHAPTER 3

SCREENING OF HIGH AND LOW RISK INDICATORS

This chapter reports on how about 1000 hypothesized indicators of high cancer risk were reduced to 62. The first three sections of the chapter present the personal, occupational, and local ambient environment indicators. The fourth section describes the creation of three factor analysis indicators. The fifth part overviews the results focusing on the presence of expected and unexpected correlations. You may wish to refer to Table 31 that includes all the correlations between the 24 cancer mortalities and 62 high risk factors.

As you review this chapter and especially Chapter 4, remember the strong cautions indicated about reading too much into the mathematical correlations. We are trying to isolate disease and place combinations which may be manifestations of specific cause/effect relationships. The scale of this study and quality of some of the data that have been used do not permit us to conclude that cause/effect relationships exist. Only detailed follow-up studies can come closer to conclusive findings.

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
		WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WF	WM	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF
1. MEDIAN INCOME	45							(297)						301											
	40							(267)						496		x 280		(268)			X(296)				
	21													599		x 772									
2. MALE MEDIAN EDUC.	45	(338)	NA		NA		NA		NA		NA		NA	NA			NA	(287)	NA		NA		NA		NA
	40	(323)	NA		NA		NA		NA		NA		NA	NA			NA	(333)	NA		NA		NA		NA
	21		NA		NA		NA		NA		NA		NA	NA		551	NA		NA		NA		NA		NA
3. FEMALE MEDIAN EDUC.	45	NA	(327)	NA		NA	X(302	NA	(452)	NA		NA		339	NA	NA		NA		NA		NA		NA	
	40	NA	(268)	NA		NA	X(291	NA	(474)	NA		NA		369	NA	NA		NA		NA		NA	255	NA	
	21	NA		NA		NA	X(443	NA	(488)	NA		NA		376	NA	NA		NA		NA		NA		NA	
4. % MALE COL LEC. +	45		NA		NA		NA		NA		NA		NA	NA			NA	(339)	NA		NA		NA		NA
	40		NA		NA		NA		NA		NA		NA	NA			NA	(392)	NA		NA		NA		NA
	21		NA		NA		NA		NA		NA		NA	NA		X 505	NA		NA		NA		NA		NA

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
		WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WF	WM	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF
5. % FEMA. COLLEGE+	45	NA		NA		NA		NA	(478)	NA	X(254)	NA		416	NA	NA		NA		NA		NA		NA	
	40	NA		NA		NA		NA	(533)	NA	X(358)	NA		403	NA	NA		NA		NA		NA		NA	
	21	NA		NA		NA		NA	(534)	NA	X(399)	NA		371	NA	NA		NA		NA		NA		NA	
6. % PROFES.	45	(271)							(464)		X(323)			345				(353)							(293)
	40	(273)							(515)		X(409)			354				(401)			(282)				
	21								(472)		X(376)			471		X 593									
7. % MANAG.	45								(424)			X 255	X 296	452								309			
	40			X 301					(290)				X 364	482								279	300		
	21								(618)											(369)					
8. % TOT. POP. POLISH	45	733	717	391		416	264	436		412		456		X 344		306	254							583	257
	40	742	771	319		441		374		454		473		X 360		438	284							590	321
	21	816	851			478		471				590		X 439		522					X(435)			641	

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	/ALL	ALL
		W M	W F	W M	W F	W M	W F	W M	W F	W M	W F	W M	W F	W F	W M	W M	W F	W M	W F	W M	W F	W M	W F	W M	W F
9. % TOT. POP. USSR	45	440	424	433	252			255		456		284		X 401		249						477		430	312
	40	363	411	273	251					462				X 335		392						540	269	280	253
	21	449	507					513					373								X(401)			412	
10. % TOT. POP. GERMAN	45								X(262)					402						X(280)		256			
	40		383									256	283	491					(319)	X(298)			271		
	21																			X(494)					
11. % TOT. POP. ITAL N.	45	434	414	514		368	443	440				483	400	X 496										573	254
	40	456	504	451		411	501	402				439	366	X 501							X(256)			574	367
	21	688	657	490		473	454	504		439		489						(370)			X(480)			679	472
12. % SAME HOUSE	45	275	296			388					257		X(275)			383									270
	40	254				476	270			275	438				X(264)	357				310				268	254
	21						402			390	500					426				398					401

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

		151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190, 191 SKIN	190, 191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
	N	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WF	WM	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF
13.% S C D A O I M U F.E N H T S. Y	45	248		322		251	303		251		290												345	326	
	40								258		264													264	
	21						535				458											403		399	593
14.% M F A E R M R A D. L.	45	NA		NA		NA	(417)	NA		NA		NA			NA	NA		NA		NA		NA		NA	
	40	NA		NA		NA	(292)	NA		NA		NA			NA	NA		NA		NA		NA		NA	
	21	NA		NA	(309)	NA	(525)	NA		NA		NA	(375)		NA	NA		NA		NA		NA		NA	(560)
15.R A C A V O W G. N E S M T. P. (15)	45	332	382	385		326	295	325		255	293	254	276	X 357										415	390
	40		326	475				256		295		344		X 568		312						410		492	332
	21	502	662	620	347	537	401	575				412		X 748		568				X(300	X(323	486		700	685
16. R A T A V E W G.X E T M I. P. (27)	45	410	423	378		377	327	348				264		X 257										446	365
	40	332	410			375																		299	
	21	540	601			474	407																	451	

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
		WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WF	WM	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF
17.R A M A V I	45	575	583	521		435	342	580		320		459		X 451										646	443
W G.C E C A	40	531	598	420		388	260	550		283		419		X 413		360								582	352
M H L P. E (28)	21	678	784	467		493		623				472		X 696		594								712	590
18.R A M A V E	45	272	253	367				352																337	
W G.T E P A	40			262				266																	
M R L P.M.S(33)	21	700	786	540		408		626				506		X 533		469								717	533
19.R A M A V E	45	410	410	476		399	302	388		266	249	263		X 376										515	400
W G.T E F A	40	292	353	416		389		290						X 356		275						305		444	
M A L P.B.S(34)	21	600	695	586		678	463	602		378				X 693		618						454		732	704
20.R A H A V I	45	353	366	505		457	284	312				291		X 388										524	316
W G.N E M E	40		395	372		442								X 320										376	
M A R P.C Y(35)	21	559	696	598		718	453	579						X 669		518						427		715	729

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190, 191 SKIN	190, 191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
		W M	W F	W M	W F	W M	W F	W M	W F	W M	W F	W M	W F	W F	W M	W M	W F	W M	W F	W M	W F	W M	W F	W M	W F
21. R A N A A V S T W G.F I E T O O M R R N P.A T(37)	45		272	301				264				307		X 360										407	
	40			251										X 286										270	
	21	531	692	482		508	309	572				421		X 736		520						402		660	676
22. E S M I P. C A (3) V G.	45	467	474	531		456	328	445		262		348		X 427										591	411
	40	396	484	469		475		375				249		X 423		284								545	477
	21	655	774	570		647	405	630		374		412		X 725		609						404		758	701
23. E S M I P. C A (7) V G.	45	416	445	480		400	310	406		254	267	311		X 392										526	398
	40	317	449	463		420		349				256		X 412		280						288		523	256
	21	696	831	590		660	440	657		365		460		X 727		606						445		791	717
24. R A A A V L W G. L E (99) M P.	45	400	427	392		362	315	363			281	250		X 331										441	387
	40	399	507	481		480	288	404				316		X 430		300						330		592	318
	21	710	813	573		648	514	660				522		X 695		551						485		820	735

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
		WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WF	WM	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF
25. AREA INDEX CONST.(15)	45	463	479	494	275	461	407	470		271	269	409	305	X 419										596	505
	40	450	573	510	281	587	485	535		262		505		X 527		278						373		687	503
	21	586	683	517		645	558	647				506		X 615								496		761	718
26. AREA INDEX TEXTL.(22)	45	505	465	426	261	454	400	446				443		X 265										587	420
	40	540	576	282		526	458	438	293			426												536	301
	21	669	688			599	592	513				552										366		664	487
27. AREA INDEX CHEMCL.(28)	45	573	569	471	262	524	428	581				508		X 401										671	443
	40	548	599	393		512	431	552	320			459		X 370										609	408
	21	631	707	443		597	499	634				531		X 566								409		740	652
28. AREA INDEX PRIMRY METALS (33)	45	554	542	479		509	371	574	276			466		X 321										657	377
	40	535	580	411		507	367	546	383			414		X 289										605	355
	21	675	726	451		605	506	648	398			596		X 493										764	632

TABLE 31

BIVARIATE CORRELATIONS; 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
		W M	W F	W M	W F	W M	W F	W M	W F	W M	W F	W M	W F	W F	W M	W M	W F	W M	W F	W M	W F	W M.	W F	W M	W F
29. AREA INDEX FABRCD METALS(30)	45	523	501	520	307	516	421	520				459	250	X 398										664	491
	40	518	577	439		573	450	521	316			413		X 380								278		627	426
	21	600	663	464		637	542	609				484		X 541						X(368)		429		730	667
30. AREA INDEX MACHN.(35) R	45	493	486	498	296	535	420	514				488		X 378										660	440
	40	463	533	407		553	436	483	297			424		X 347										590	386
	21	577	642	441		622	532	592				492		X 500						X(369)		404		706	641
31. AREA INDEX TRANS.(37) E	45	425	466	462	265	476	381	489				500		X 406										627	396
	40	372	473	380		455	371	440	263			440		X 373										557	354
	21	556	633	400		549	49[illegible]	567				516		X 526								372		673	612
32. AREA INDEX IC (3)	45	543	532	509	296	539	434	552				490		X 402										682	470
	40	516	577	415		549	44[illegible]	526	318			439		X 371										614	410
	21	607	676	450		620	525	616				507		X 539								415		729	655

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

		151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
	N	W M	W F	W M	W F	W M	W F	W M	W F	W M	W F	W M	W F	W F	W M	W M	W F	W M	W F	W M	W F	W M	W F	W M	W F
33. AREA INDEX SIC (7)	45	546	542	524	300	538	445	551				501	259	X 411										689	493
	40	518	592	431		564	457	534	316			468		X 384								259		639	420
	21	632	702	454		634	551	628				539		X 533								424		749	662
34. AREA INDEX ALL (99)	45	552	552	491	270	527	456	534		240		467	273	X 383										651	495
	40	521	574	406		552	469	525	337			485		X 337										629	405
	21	619	664	431		612	571	617				581		X 462								398		740	626
35. DISTA- NCE CONST.(15)	45	569	592	513	282	566	521	611		297		592	333	X 550								265		706	495
	40	522	600	418	285	548	529	571	248	252		567		X 524		264						263		650	468
	21	636	728	427		639	545	612				588		X 617		389								733	595
36. DISTA- NCE EXTS.(22)	45	537	529	481	302	542	497	582		261		538	263	X 454										659	452
	40	494	555	354	260	539	516	546	248			480		X 418										571	393
	21	631	728	380		604	590	589				543		X 504		390								707	566

Reproduced from best available copy.

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L.INTEST.	153 L.INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
		WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WF	WM	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF
37.D A C I N H S C E T-E M.(28)	45	507	591	457	278	547	483	613		293		552	268	X 498										657	449
	40	559	610	372	289	529	490	582	264	261		513		X 464		306								601	437
	21	660	765	403		637	476	593				528		X 678		512								705	574
38.D A P H I N R E S C H T T- E Y.A L (33) S	45	517	502	463		521	417	576				451		X 428										611	358
	40	477	516	355	253	513	398	528	282			385		X 388		274								542	341
	21	657	743	396		655	476	555				517		X 657		550								689	561
39.D A F H I N A E S C B.T T-E A L (34) S	45	565	561	521	300	571	497	604		282		560	291	X 517										702	470
	40	526	587	409	286	566	506	565	264			502		X 487		279								628	433
	21	638	727	426		665	545	599				540		X 637		453								723	593
40.D A M I N A S C C T-E H. (35)	45	536	546	510	290	573	485	575				551	273	X 531										682	443
	40	500	582	396	274	573	499	530				480		X 503		267								604	413
	21	628	728	414		672	539	590				518		X 654		475								708	592

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
		WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WF	WM	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF
41. DIST. TRANS. (37)	45	477	506	508	275	512	469	584				566	317	X 567										674	429
	40	427	526	398	283	498	473	536				506		X 544		259								605	418
	21	621	718	407		623	534	588				535		X 687		481						379		717	613
42. AVG. DIST. SIC (3)	45	570	573	501	293	570	494	605		276		561	280	X 521										688	459
	40	535	600	395	286	560	503	565	255			504		X 489		288								617	432
	21	645	744	416		660	521	596				534		X 659		483								715	588
43. AVG. DIS. SIC (7)	45	564	573	508	290	567	504	610		279		570	297	X 528										695	468
	40	522	592	401	286	555	511	569	254			520		X 496		277								625	438
	21	644	721	417		658	539	601				554		X 645		453								723	592
44. DIST. ALL (99)	45	585	606	492	264	579	518	610		285		566	299	X 518										691	478
	40	542	610	402	268	561	523	571	261			541		X 485		256								637	449
	21	638	730	415		647	546	607				576		X 604		383								725	588

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH WM	151 STOMACH WF	153 L. INTEST. WM	153 L. INTEST. WF	154 RECTUM WM	154 RECTUM WF	155 LIVER WM	155 LIVER WF	157 PANCREAS WM	157 PANCREAS WF	162 LUNG WM	162 LUNG WF	170 BREAST WF	177 PROSTATE WM	180 KIDNEY WM	180 KIDNEY WF	181 BLADDER WM	181 BLADDER WF	190,191 SKIN WM	190,191 SKIN WF	204 LEUKEMIA WM	204 LEUKEMIA WF	ALL WM	ALL WF
45, % CHEM. (28)	45					X(248												409		573					
	40																	456	299	616					
	21														X381			541		702					
46.% TRANS. (31)	45											308													
	40											284													
	21															535									
47. % AGRIC.	45	(383)	(344)	(339)		(333)	(339)	(265)						X 378							252			(317)	
	40	(344)	(309)	(322)		(316)	(305)			(259)				X 405							520			(297)	
	21	(610)	(603)	(754)		(400)		(639)		(594)		(600)		X 567		(470)		X 348						(740)	(422)
48. TSP	45	301	361	383				321			293	330	289	X 303										420	328
	40			293								420				261								350	
	21	680	717	525				561				603		X 544		523								709	506

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH WM	151 STOMACH WF	153 L. INTEST. WM	153 L. INTEST. WF	154 RECTUM WM	154 RECTUM WF	155 LIVER WM	155 LIVER WF	157 PANCREAS WM	157 PANCREAS WF	162 LUNG WM	162 LUNG WF	170 BREAST WF	177 PROSTATE WM	180 KIDNEY WM	180 KIDNEY WF	181 BLADDER WM	181 BLADDER WF	190,1 SKIN WM	190,1 SKIN WF	204 LEUKEMIA WM	204 LEUKEMIA WF	ALL WM	ALL WF
49. TSP/AREA	45	400	512	503	249	469	392	495		255	248	504	301	X 384										641	451
	40	428	531	423		480	359	467	254			567		X 342										631	339
	21	685	717	483		596	527	665	379			660		X 495										800	645
50. TSP/WIND	45	469	488	539	281	415	370	507		308	295	520	359	X 449										654	494
	40	409	527	512	301	412	328	522		309		645	291	X 492		344						297		708	463
	21	744	785	550		534	422	692				691		X 628		445								842	669
51. TONS COAL EQUIV. 1930-72	45	532	537	408		344	330	485			265	456		X .314										550	348
	40	484	506	313		250		450	261			519											X(261)	514	
	21	662	673	425		422		640	508			695												712	466
52. TONS COAL EQUIV. AREA, V.	45	493	471	322		423	381	464	269			490												541	306
	40	463	461	273		395	379	435	347			486												505	270
	21	549	533			465	489	531	437			639												627	436

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
		WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WF	WM	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF
53. T C E W O O Q I N A U N S L I D/D V. I C	45	597	504	448		482	436	569	277			569		K 345										666	418
	40	547	563	361		436	410	525	359			567		K 275										609	342
	21	636	641	384		496	472	615	460			706												713	501
54. V H T	45		308	320						267	341			K 393										318	
	40													K 394								257			
	21	438	526	485				416				437		K 642		680								508	370
55. V D S H E I T/ N- T Y	45	421	466	472		396	359	400		340	346	385	318	K 509								283		557	460
	40	316	449	445		378	308	324		377	298	450		K 665		363						412		589	412
	21	648	741	524		558	419	609				517		K 727		597						372		758	705
56. V W H I T N D/D I S	45	401	466	494		398	342	365		360	371	392	316	K 573								340		569	449
	40	270	423	432		351	256			368	310	389		K 678		328						454		544	389
	21	690	794	550		541	384	653				602		K 767		640								779	658

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L.INTEST.	153 L.INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
		WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WF	WM	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF
57.% SELF SERVCE WATER	45	(339)	(371	(412)		(344)	(397)	(384)				(401)												(500)	
	40	(269)	(384	(297)	(260)	(405)	(383)	(258)				(420)					(267)							(486)	(307
	21	(577)	(689	(416)	(423)		(487)	(552)				(501)		X(563)		(524)								(639)	(500
58.% ALL SURFCE WATER	45	398	425	318		410		304									318							386	
	40	323	390			393											335							308	
	21	703	796			621	302	478				388		X 408		405								546	
59. AVG. POP DNSTY	45	495	509	484	264	483	435	492		265	263	427		X 404										606	497
	40	482	566	465	268	571	500	540	271			528		X 440								307		672	455
	21	601	664	481		632	584	641				575		X(521								440		762	666
60. P I C O T O L/A R L. R	45	502	479	271		443	396	364				328												454	
	21																								

Reproduced from best available copy.

TABLE 31

BIVARIATE CORRELATIONS: 24 BODY SITES AND 62 HIGH RISK INDICATORS

	N	151 STOMACH	151 STOMACH	153 L. INTEST.	153 L. INTEST.	154 RECTUM	154 RECTUM	155 LIVER	155 LIVER	157 PANCREAS	157 PANCREAS	162 LUNG	162 LUNG	170 BREAST	177 PROSTATE	180 KIDNEY	180 KIDNEY	181 BLADDER	181 BLADDER	190,191 SKIN	190,191 SKIN	204 LEUKEMIA	204 LEUKEMIA	ALL	ALL
		WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF	WF	WM	WM	WF	WM	WF	WM	WF	WM	WF	WM	WF
61.U I C K/W N C B D. I A U E N S.	45	507	537	502	317	405	358	488		329	265	523	383	X'512								318		633	530
62.S E .. S O C E T C O I A I N C T O/O. U S	45	(346)							(459)					410				(328)							

NOTES: WM is white male; WF is white female

NA is correlation not appropriate;

( ) is a negative correlation;

X before a correlation signifies an assumed spurious correlation.

Personal Environment Risk Factors

Twenty-three indicators of elevated personal risk were defined in Chapter 1. Fourteen of the 23 were retained (Table 32). An explanation of each decision is given below.

TABLE 32

Screening Of Personal Environment Factors

	Indicator	Retained
1.	Median income, family and unrelated individuals	Yes
2.	Median school years completed, males (≥ 25 years old)	Yes
3.	Median schools years completed, females (≥ 25 years old)	Yes
4.	% male high school and 1-3 years of college completed	No
5.	% female high school and 1-3 years of college completed	No
6.	% male at least college educated	Yes
7.	% female at least college educated	Yes
8.	% male less than 4 years high school	No
9.	% female less than 4 years high school	No
10.	% professional and technical employment	Yes
11.	% managerial employment	Yes
12.	% Polish Foreign stock	Yes
13.	% U.S.S.R. Foreign stock	Yes
14.	% German Foreign stock	Yes
15.	% Italian Foreign stock	Yes
16.	% same house as five years earlier	Yes
17.	% different house in same county as five years earlier	Yes
18.	% male single	No
19.	% female single	No
20.	% male married	No
21.	% female married	Yes
22.	Fertility ratio	No
23.	Persons per household	No

The income variable was retained. It consistently produced the expected positive association with female breast cancer. Median income also correlated negatively, though not consistently, with liver and bladder cancer. Inconsistent and unexpected correlations were produced between income and kidney cancer (positive) and skin (negative). The strongest associations of income were with female breast cancer.

The eight educational achievement variables correlated as expected. As indicators of socioeconomic status, we presumed that they would correlate with lung, stomach, liver, and bladder cancer mortality. The expected relationships with female breast, bladder, stomach, liver and leukemia (very inconsistent) cancer were found. One unexpected correlation was found: female at least college completed and female pancreas cancer. We have no explanation for the correlation and can add the confirmation that percent professional and managerial employment were also associated with female pancreas cancer. This association may be but is not necessarily spurious. It will be discussed in Chapter 4.

Overall, all eight education variables produced the expected correlations. Four were retained and four were eliminated as redundant. The most general (median school years completed) and highest correlations (completed college) were retained.

The four ethnic variables were retained. The Polish variable, as expected, demonstrated the strongest correlations with stomach and lung cancer, although the expected association with female lung cancer did not materialize. The stomach cancer association is extremely strong. The U.S.S.R. population also generally followed the expected pattern. High correlations were found with leukemia and stomach. The expected associations with cancer of the large intestine, pancreas, and kidney also materialized; however, they were inconsistent. An expected strong association between cancer of the lung was found to be weak and inconsistent. Finally, an expected correlation with cancer of the rectum did not appear. Overall, the U.S.S.R. group, a surrogate for Jewish, demonstrated the expected correlations, but the correlations are not as strong as we had anticipated.

Whereas the U.S.S.R. correlation was lower than we had expected, the Italian correlations were stronger than anticipated. Briefly, given the lack of clarity in the literature about rates among the Italian ethnic population, we formulated conservatively low hypotheses. The correlations turned out to be slightly stronger than those exhibited by the U.S.S.R. ethnic population and not quite as strong as those of the Polish group.

We expected that the German variables would produce the weakest associations because the population is split between Jewish and non-Jewish elements and because the literature does not report unusual rates for the German population. The German population correlations were as expected lower than those for the other three ethnic groups and they were not consistent. The expected correlations with cancer of the stomach, leukemia, and lung were found.

In conclusion, all four of the ethnic variables were retained. The Polish and the German population indicators, as expected, ranked respectively, first and fourth in association with cancer mortality. However, the USSR and Italian indicators switched ranks. The USSR variable, in general, was third instead of second in association with cancer. The Italian variable was expected to be third; it turned out to be second. To reiterate, these foreign stock indicators are considered to be surrogates for cultural factors including diet, smoking, drinking, medical care and others.

Chapter 4 will explore the possible reasons for the reversal of the U.S.S.R. and Italian group results in the context of specific body sites. At this point, the reasons will briefly be reviewed. First the inconsistent literature findings regarding the Italian ethnic group may have misled us into developing hypotheses that were too conservative. Second, and related to the first reason, the Italian and U.S.S.R. (Jewish) populations in our study area may be affected by occupational and local ambient environmental factors differently than these two ethnic groups in the United Staes as a whole. Specifically, in comparison to the U.S.S.R. and German populations, the Italian and Polish populations live in areas that generally have higher risk occupational and local ambient environments (Table 33).

TABLE 33

Comparison of Occupational Mix and Local Ambient Environments of 4 Ethnic Groups (45 place study area) Correlation Values

High Risk Indicator Ethnic Group	All industry gravity-potential model	Tons coal equivalent wind direction/ distance indicator	Air Emissions/ wind direction/ distance indicator	% Population served by Public potable surface water
% Polish	.592	.525	.427	.437
% Italian	.509	.441	.361	.388
% U.S.S.R.	.360	.314	.522	.156
% German	.128	.042	-.065	-.131

Whereas the preponderance of the ethnic group correlations followed the expected pattern, two groups of significant correlations were unexpected. Inconsistent, negative associations were found between three of the ethnic groups and male skin cancer mortality. These correlations were unexpected; but perhaps they are real. We have no way judging because we were unable to find literature dealing with skin cancer mortality among the different ethnic populations.

The most surprising correlations were between all four ethnic groups and female breast cancer mortality. We had expected to find negative associations with the Polish and Italian populations, no association with the U.S.S.R. population and no association or a slight positive association with the German population. The surprising result was postive correlations with all four groups. The correlations were stubbornly consistent. Furthermore, they varied little among the four populations. In the case of the 45 place study area, the highest correlation was .496 with proportion Italian and the lowest was .344 with proportion Polish. In no other case did all four ethnic groups associate the same way with cancer of a body site. While the correlation values are not the highest found in the study, they are statistically significant and so consistent as to warrant strong attention. Chapter 4 will review the extremely interesting patterns of white female breast cancer in detail.

The two stability variables (% same county, different house; % same house) were expected to demonstrate low positive correlations with many of the cancers. They did and were retained.

The four life style variables were expected to demonstrate that married people had a lower cancer rate and unmarried people higher rates. The results were mixed. The male single and married variables demonstrated the expected associations for only 1970. For example, the variable proportion male married in 1970 yielded correlations ranging from -.248 to -.484 with eight of the twelve male body sites. However, the 1950 correlations were insignificant, leading us to conclude that if there are any associations, they are probably related to internal migration. Following the time consistency guideline, the two male life style variables were eliminated.

The female married variable were retained because a strong and consistent correlation was found with cancer mortality of the rectum and with the general urban-industrial cancer corridor factor.

The final two personal environment variables were the surrogates for early age at childbirth and large families. We assumed that these would be inversely associated with female breast cancer. The 1970 fertility rate produced a low negative correlation. However, the 1960 fertility rates were insignificantly correlated. Furthermore, a host of unexpected seemingly spurious correlations between these variables and cancers were produced. Accordingly, both variables were dropped.

Summarizing, 14 of the 23 personal environment variables passed the screening. They lead to some of the most interesting results that will be discussed in detail in Chapter 4.

Occupational Environment Risk Factors

More than 700 occupational variables were gathered and created. Two of the six sets were rejected on the grounds of incomplete theory and for confusing results: the location quotient; and the employment/population indices. The location quotients were rejected because of inconsistent results. Few of the expected positive associations were found. Those that were observed appeared in one year, disappeared in another, and sometimes reappeared in a third, only to disappear again in the fourth time period. Furthermore, some negative correlations were found. Overall, the location quotients added little and were therefore eliminated from further consideration.

The employment/population indicator was rejected on the grounds of inconsistent and confusing results. Each initial test of the index contained 576 bivariate correlations (24 employment/population x 24 cancer body sites.) The 1947 test matrix produced 88 significant correlations, the 1958 data 90, the 1967, 9, and the 1972 only 5. Twenty-nine significant correlations could have been expected by chance. In short, significant results were found in 1947 and 1958 but not especially in 1967 and 1972. Furthermore, whereas we had expected production employment to produce the strongest correlations and establishments the lowest correlations, the majority of the significant correlations were between cancer and the ratio of establishments to population. For example, 51 of the 90 significant correlations found with the 1958 data were between establishment/population ratios and cancer. Many of the correlations such as chemicals with cancer of the liver, lung, and stomach; textiles with female stomach, rectum, and liver were expected. However, most of the expected correlations did not appear. When combined with the disappearance of the significant correlations after 1958, we chose to eliminate the employment/population index from further consideration.

There is, however, an important implication to be drawn from the population index. Specifically, whereas the employment/population index emphasizes large facilities, the establishment/population index emphasizes small facilities. An important implication of the correlation between the establishment/population index and cancer is that research on the hazards of industry should not be confined to the largest facilities. Smaller facilities may be high risk sources of carcinogenic substances.

Screening of the occupational indicators resulted in the selection of four types for the final analyses. Three of the categories are occupation by workplace, one is occupation by residence. These four types of indicators include thirty-three specific variables (Table 34). Each variable consistently appeared in each time period.

TABLE 34

Occupational Variables Retained for Final Correlations

SIC Code & Category	Workplace Employment[a]	Workplace Employment/ Land Area[a]	Residence % Employment[b]	Workplace Gravity Potential[a]
1. 15 - Construction	Yes	Yes	No	Yes
2. 22 - Textiles	Yes	Yes	No	Yes
3 28 - Chemicals	Yes	Yes	Yes	Yes
4. 33 - Prim. Met.	Yes	Yes	No	Yes
5. 34 - Fab. Met.	Yes	Yes	No	Yes
6. 35 - Mach.	Yes	Yes	No	Yes
7. 37 - Transp.	Yes	Yes	Yes	Yes
8. Agricul.	No	No	Yes	No
9. SIC 28, 34, & 35	Yes	Yes	No	Yes
10. SIC's 15, 22, 28, 33, 34, 35, 37	Yes	Yes	No	Yes
11. All manufacturing	Yes	Yes	No	Yes

a - average of 1947, 1958, 1967, and 1972.
b - average of 1950, 1960, and 1970.

In addition to temporal and geographical consistency, these variables had to pass other tests. First, we hypothesized in Chapter 1 that the production employment data would produce higher correlations than total employment data, that both employment indicators would yield higher correlations than the number of establishments, and that the earlier years should produce stronger results than the later years.

The production employment versus total employment test was not made because production and total employment are strongly associated in the region. The correlations between manufacturing production employment and total manufacturing employment in the region in 1947, 1958, 1967 and 1972 are .999, .987, .999, and .998 respectively. The employment variables did not always produce higher correlations than the establishment variables. If important variables are considered to be those that were significant at the .05 level, then we cannot conclude that the employment data yielded higher correlations than the establishment data. In a little more than half of the cases, when one of the correlations was significant, the employment association was higher than the establishment correlation. If, however, a correlation value of .50 or greater is considered important, then employment is, as expected, a better indicator of high risk than is establishments. Specifically, 88 percent of the correlations of at least 0.50 were with employment data.

The last initial test was the assumption that the earliest time period would produce stronger associations than the latest. Using employment in 1947 and 1972 as the indicator variables, we found the hypothesis to be verified. In 72 percent of the cases when a statistically significant (.05 level) correlation was found, the 1947 employment distribution produced a higher correlation than the 1972 intraregional distribution of employment. Furthermore, about twice as many correlations of at least 0.5 were produced by the 1947 data than resulted from the 1972 information.

Summarizing, the use of occupational indicators is supported by consistency tests (correlation reappears in multiple time periods and in subregions of the study area), and by control variable tests (employment stronger correlation than establishment and earlier time period stronger correlations than later time period).

Upon initial verification of the occupational variables, they were transformed to averages. The 1947, 1958, 1967, and 1972 occupations by place of work were divided by 4 to yield 1947-1972 average employment vectors. The 1950, 1960 and 1970 occupation by place of residence matrix was divided by 3 to produce 1950-1970 average employment vectors.

These average vectors were used to test the final two control hypotheses: (1) the manufacturing/land area index will demonstrate higher correlations than the untransformed manufacturing data; and (2) the manufacturing/ distance index will demonstrate higher correlations that either the manufacturing/ land area index or the untransformed manufacturing data. Using the average employment vector for the 45 place study area, we found that the hypotheses were verified. In 88 percent of the cases when one or both demonstrated significant correlations, the manufacturing/ land area index exhibited higher correlations with cancer mortality than the untransformed manufacturing data. Likewise, in 85 percent of the cases, the manufacturing/distance index demonstrated higher associations than the manufacturing/ land area index when at least one was significantly correlated with a cancer category.

Over, thirty-three occupation variables joined the fourteen personal environment variables in the final analysis stage.

Local Ambient Environment Risk Factors

The air quality and emissions results were contradictory. The air quality data were eliminated, while the air emissions surrogates demonstrated strong associations. It was suggested in Chapter 1 that the air quality data were a dubious representation of air quality in many of the study area's counties. If judged by the expected associations between air quality and cancer, our misgivings about the data were warranted. Neither the 1953-1957 nor the 1972-1975 air quality variables correlated positively with lung cancer. When the region was reduced in size from 45 units to the 21 New Jersey counties, a low positive association between white male lung cancer and micrograms of suspended particulates resulted. One weak correlation does not justify the retention of a variable. Both variables exhibited positive correlations with stomach, large intestine, rectum, liver, and all cancers. However, the correlations were inconsistent. The two air quality variables were eliminated, despite the fact that some expected correlations were produced, because the results were inconsistent and the expected association with lung cancer did not materialize.

Whereas the air quality parameters were rejected, the air emissions surrogates were retained. The air emissions indicator consistently correlated with male lung cancer and also correlated with female lung cancer and with other body sites. The strongest correlations between total suspended particulates (TSP) emissions and male white lung cancer are in the New Jersey portion of the region.

When TSP emissions were divided by land area, the expected positive associations increased. When the TSP emissions/wind direction/distance model was used most of the expected associations increased again.

Having successfully used the wind direction/distance model, we chose to further test it. If the model is a reasonable abstraction of the movement of pollutants in the study area and the correlations with cancer mortality are not spurious, then the correlations should change if we artificially change the wind direction. If the wind blew from the east 100 percent of time, we would expect the correlations to decrease because Philadelphia's emissions would go west instead of predominantly north and east. New York City's emissions would affect northern New Jersey, rather than places to the north and east. The Atlantic Coast and central New Jersey counties would be the beneficiaries of winds that always blow from the west.

Whereas we would expect the correlations to decrease if the winds always came from the east, we expected that they would remain about the same if the winds always came from the southwest. Southwesterly winds would blow and concentrate the emissions along the urban-industrial corridor. Finally, we expected a small decrease in the correlations if the wind always came from the west.

The results for the three most important categories of air pollution related to cancer are listed below (Table 35). We expected that the southwesterly wind pattern would produce about the same correlations. The results generally support the hypotheses. Four of the six correlations are higher than the actual pattern; the increases, however, are not much stronger. Next, the westerly wind pattern should have produced lower correlations than the actual winds. In five of the six cases, the expected materialized. However, the westerly correlations were higher than we had anticipated.

Finally, as expected the correlations are significantly lower when the winds are made to come from the east 100 percent of the time. Indeed, the lung cancer correlations are insignificant.

TABLE 35

Test of Air Emissions/Wind Direction/Distance Indicators Correlations with Cancer Mortality

Wind Direction	Lung		Stomach		All	
	Male	Female	Male	Female	Male	Female
Actual pattern	.520	.359	.469	.488	.654	.494
From Southwest 100% of time	.493	.364	.570	.533	.670	.458
From West 100% of time	.512	.399	.327	.365	.573	.435
From East 100% of time	-	-	.360	.395	.357	.351

In conclusion, the tests of the TWP/wind direction/distance model support the contentions that in so far as the movement of air emissions is concerned, this crude model is a reasonable abstraction from reality. While we are satisfied with the wind direction/distance model, our confidence would be greatly increased if the results could be replicated with a combination of data on the emissions of carcinogenic substances and an air quality diffusion model. At this point, the model is suggestive and speculative about what might be happening.

The successful testing of the air emissions indicators suggests that our efforts to collect fossil fuel and motor vehicle data were unnecessary. While hindsight suggests that these two sets of indicators were not required, they were tested. The twenty-four initial fossil fuel variables were reduced to three: total tons coal equivalents burned (TCE), 1938-1972; TCE, 1938-1972/land area; and TCE, 1938-1972/wind direction/distance. The coal variables (1938-1947; 1950-1965; 1970-1972; and 1938-1972) did not consistently produce significant correlations with lung cancer and exhibited lower correlations with the other body sites than the more inclusive tons coal equivalent (TCE) variables. The highest TCE and cancer correlations were with the time period 1950-1965. However, the 1950-1965 and 1938-1972 correlations were so close that we applied the fallback rule (see Chapter 1) and adopted the more general indicator.

The final three TCE, 1938-1972 variables produced the expected pattern of correlations. The TCE/wind direction/distance variable exhibited the highest correlations.

Although lower than expected, the vehicle miles traveled (VMT) variable also demonstrated the expected associations. Therefore, three VMT variables were retained: (1) VMT; (2) VMT/land area; (3) VMT/wind direction/distance. Overall, nine indicators of air pollution were added to the fourteen personal environment and thirty-three occupational environment variables.

The final two retained local ambient environment indicators are (1) % of county population self-service water in 1963; and (2) % of county population served public potable surface water in 1963. Counties that had a high proportion of their population served by public potable surface water in 1963 had higher rates of digestive and urinary tract cancer. The converse was true of counties that had a high proportion of self-serviced residents.

The fifty-ninth variable is the control variable average population density, 1950-1970. As indicated in Chapter 1, population density is a surrogate for all the high risk urbanization and industrialization factors that are associated with cancer. If an indicator produces a higher correlation with cancer than does population density, we assume that the high risk indicator has a general and perhaps an unique association with cancer.

Factor Analysis Indicators

As the statistical results were examined, many of the high risk factors were to be found in the same counties. Some counties had high risk personal, occupational and local ambient environmental factors. Other counties low risk combinations. For example, among the forty-five counties, Essex County, New Jersey ranked third in the TSP/wind direction/distance index, eighth in percentage U.S.S.R., and sixth in the number of chemical workers. Dutchess County, New York ranked fourtieth in chemical workers, twenty-seventh in percentage U.S.S.R., and thirty-fifth in TSP/wind direction/distance. The Essex cancer rates should be much higher than the Dutchess rates because the former has multiple high risks, the latter multiple low risk factors. The correlations among percent U.S.S.R., TSP/wind direction/distance and chemical workers by place of occupation are as follows:

	%U.S.S.R.	TSP/wind direction distance	Chemical Workers
% U.S.S.R.	1.000	.522	.465
TSP/wind direction/distance	.522	1.000	.860
Chemical workers	.465	.860	1.000

These correlations indicate that it is extremely difficult to statistically separate the interaction impacts of these high risk factors. Indeed, separating them may be artificial and not a representation of their complex interactions in the real world. A set of factor analyses were made to develop statistically uncorrelated variables that can include the interactions of personal, occupational, and local ambient variables. Three important factors emerged from the analysis. They are called urban/industrial core, Polish/Italian corridor and socioeconomic status. The three factors accounted for between 59 and 76 percent of the variance in the high risk factor matrices. Two other factors were produced. However, they significantly correlated only once with cancer mortality and were inconsistent between different factor rotations. Accordingly, while of interest, they are not relevant and are therefore not reviewed. A series of preliminary factor analysis runs were made to eliminate statistically redundant high and low risk indicators. These preliminary runs reduced the number of high and low risk indicators from 59 to 23. Using the 23 variable matrix, the first factor is urban/industrial core

(Table 36). It describes counties whose populations live in high density residential settlements surrounded by manufacturing facilities, highways, and electricity generating stations.

TABLE 36

Urban/Industrial Core Factor

	Variable	Factor Loading
1.	Population Density	.928
2.	TSP/wind direction/distance index	.921
3.	VMT/wind direction distance index	.832
4.	TCE/wind direction/distance index	.775
5.	All manufacturing employed by workplace/distance index	.770
6.	% Russian foreign stock population	.630
7.	% female married	-.541
8.	% Polish foreign stock population	.437
9.	% Italian foreign stock population	.338
10.	% surface water	.306

Secondary characteristics of these counties are their relative concentrations of unmarried females and Russian, Polish, and Italian ethnic populations.

The factor scores clearly identify these places as New York City (4.69), Philadelphia (2.37), Hudson (2.19), and Essex Counties (1.05) (Table 37). (Recall that 0.00 is a neutral score; 1.00 a high score). These entities contain the four major central cities of the region: New York, Philadelphia, Newark, and Jersey City. The six counties that most closely identify with the urban/industrial core factor comprise the six highest counties in male white cancer rate and the urban-industrial cancer corridor group. The initial correspondence between the urban/industrial core and urban-industrial cancer corridor factors is startling given the fact that they were arrived at with different variables.

At the other end of the spectrum are less urbanized/industrialized counties. They have relatively few urbanized and industrialized areas, few Russian nationals and their descendants, and most of their female population is married. The counties strongly manifesting these characteristics are predominantly in Pennsylvania (Table 37).

TABLE 37

Counties Most Strongly Identifying With The Urban/Industrial Core Factor

County	Factor Score	Rank	Rank in white male Cancer rate, 1950-1969
1. New York City	4.69	1	4
2. Philadelphia	2.37	2	2
3. Hudson	2.19	3	1
4. Essex	1.05	4	5
5. Nassau	.87	5	6
6. Middlesex	.59	6	3
7. Carbon	-1.23	45	39
8. Northampton	-1.11	44	27
9. Lehigh	- .72	43	38
10. Warren	- .71	42	23
11. Litchfield	- .67	41	31

The counties demonstrating characteristics most opposite to the urban-industrial core have among the lowest cancer rates, but not the lowest. Overall, the urban/industrial core factor should imperfectly fit the urban-industrial cancer corridor group.

The second statistically derived variable is the Polish/Italian corridor factor. It describes urban corridor counties that have a large proportion of Polish and Italian citizens, the incomes of the people in these counties are relatively low (lower middle income) and they tend to be relatively stable in terms of population mobility. Manufacturing and electricity generation are common in these counties, but not as common as in the urban/industrial core counties (Table 38).

TABLE 38

Italian/Polish Corridor Factor

Variable	Factor Loading
1. % Polish foreign stock population	.645
2. % Italian foreign stock population	.616
3. % self service water	-.555
4. % Agricultural workers	-.489
5. All employed manufacturing/wind direction/distance index	.484
6. Median income	-.460
7. % surface public potable water	.432
8. % same house	.363
9. Tons coal equivalent consumed/wind direction/distance index	.361

May we conclude that the relatively stable and lower middle income people in these counties who work in and live near manufacturing facilities are Polish and Italians? While recognizing the fallacy of misattributing characteristics to individuals based on county data, the above description is consistent with the literature. As Chapter 4 will suggest, epidemiological studies of these two populations in the counties that most closely model the factor could be extremely revealing.

The word corridor is used to help describe the factor because the counties that most closely identify with its factor loadings lie along the urban-industrial corridor (Table 39). Six of seven are in New Jersey. Two (Hudson and Middlesex) identified strongly with both the high risk factors. Among the seven counties, only Somerset does not have a white male cancer rate of greater than 200/100,000. New York City, Philadelphia, Essex, and Nassau Counties closely identified with the urban/industrial core factor, but not with the Polish/Italian corridor factor.

TABLE 39

Counties Most Strongly Identifying with the Polish-Italian Corridor County Factor

County	Factor Score	Rank in Polish/ Italian Corridor Factor	Rank in White Male Cancer Rate
1. Hudson	2.47	1	1
2. Passaic	2.19	2	7
3. Bergen	1.72	3	12
4. Union	1.54	4	11
5. Middlesex	1.40	5	3
6. New Haven	1.15	6	10
7. Somerset	1.01	7	32
45. New Castle	-1.70	45	21
44. Monroe	-1.62	44	33
43. Sullivan	-1.39	43	36
42. Cape May	-1.39	42	18
41. Chester	-1.10	41	45
40. Rockland	-1.07	40	38
39. Putnam	-1.07	39	17

Those counties that have the converse of the characteristics attributable to the Polish/Italian corridor counties tend to have lower cancer rates. However, there is no overlap among these seven low risk counties and the five low risk counties that contrasted sharply with the urban/industrial core counties.

The third factor distinguishes high and low socioeconomic status counties (Table 40). The high status counties (Table 41) have many well-educated, well paid people who are engaged in professional occupations. We would expect these counties to have high rates of female breast cancer, perhaps, leukemia, and low rates of lung, stomach, liver, and bladder cancer. The low socioeconomic status counties should demonstrate the opposite pattern. Low socioeconomic status should mark a county as high risk for males. For females decreased risk for breast cancer is balanced by higher risk for liver, bladder and other sites.

Overall, the three high risk factor analysis variables should be useful in multivariate statistical analyses that are marked by interacting high risk factors. The factors are not correlated.

While the factors are not correlated, a few counties can strongly identify with more than one factor (Figure 1). Hudson County strongly identifies with all three factors of high risk to males and Middlesex and Philadelphia identify with two of the three. These three counties have the highest white male cancer mortality rates in the study area, again a startling correspondence. Beyond these three counties, simple generalizations are more difficult to justify. For example, Chester County has the lowest white male cancer rates; it identifies strongly only with one of the three low risk factors. The two other counties with white male cancer mortality rates below 175 are Dutchess and Bucks. Dutchess and Bucks do not strongly identify with any of the three low risk factors.

TABLE 40

Socioeconomic Status Factor

Variable	Factor Loading
1. % male college plus	.974
2. % male median education	.954
3. % professional employment	.924
4. % female median education	.919
5. % female college plus	.890
6. Median income	.767
7. % managerial workers	.578

TABLE 41

Counties Strongly Identifying with the
Socioeconomic Status Factor

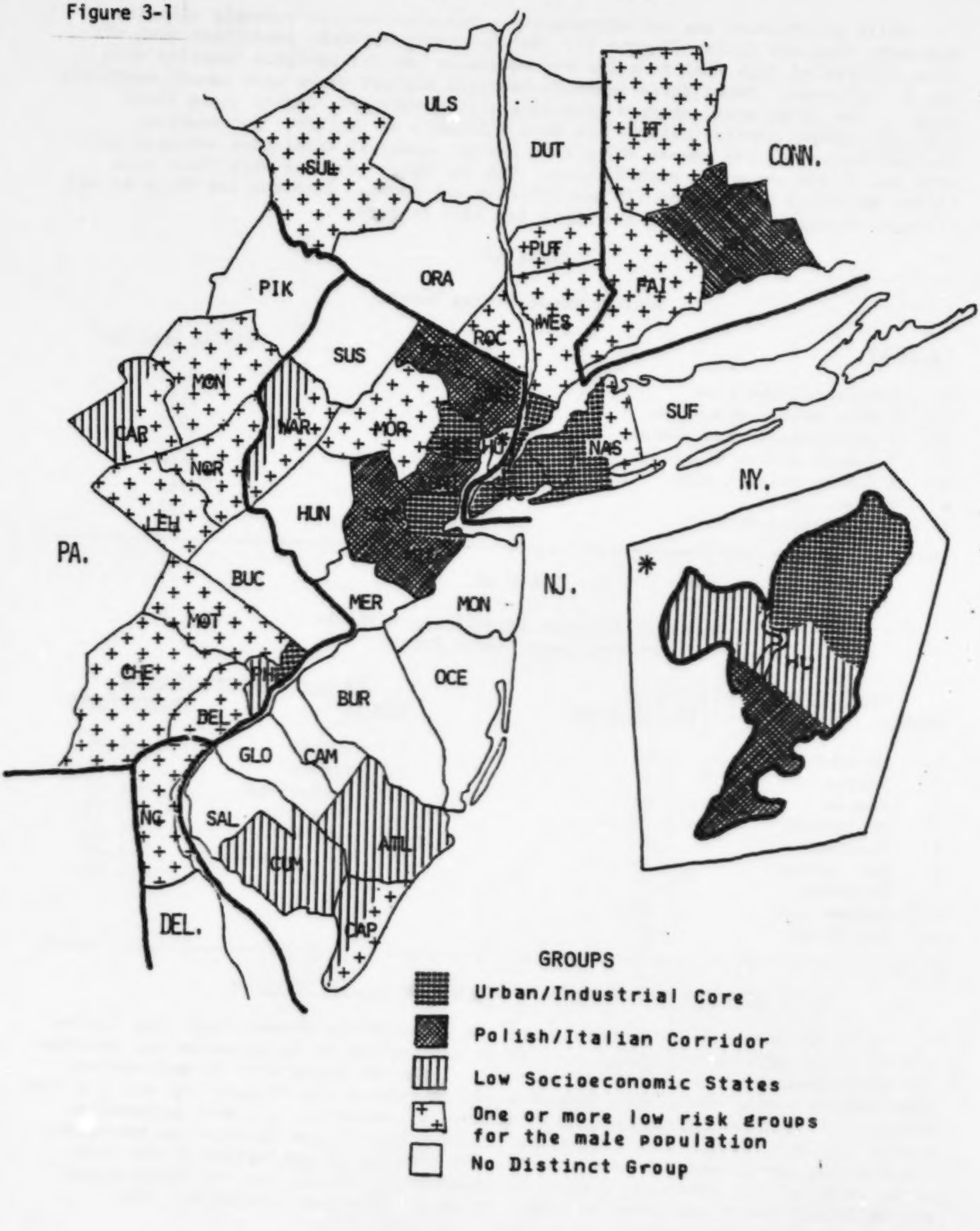
High Socioeconomic Status		Low Socioeconomic Status	
County	Factor Score	County	Factor Score
1. Westchester	2.34	1. Hudson	-1.86
2. Morris	1.59	2. Carbon	-1.74
3. Nassau	1.56	3. Cumberland	-1.41
4. Montgomery	1.50	4. Atlantic	-1.15
5. Rockland	1.41	5. Cape May	-1.11
6. New Castle	1.29	6. Philadelphia	-1.05
7. Delaware	1.25	7. Warren	-1.00
8. Putnam	1.10		
9. Fairfield	1.10		

Expected and Unexpected Correlations: An Overview

Prior to presenting the analysis of the associations between high risk factors and each type of cancer in Chapter 4, it is appropriate to recapitulate and overview the effectiveness of the screening process. The study began with 24 age/race/sex standardized cancer mortalities and about 1000 gathered and created high and low risk indicators. Many of the indicators were somewhat redundant, but were gathered to ensure that the results were consistent. By setting minimum correlation standards, checking for consistency of associations over time and in subregions of the study area, using control diseases, control independent variables, control correlations, and employing factor analysis, we sought to select the best indicators. The initial variables were reduced to 62 high risk indicators.

COMPOSITE RISK GROUPS

Figure 3-1


ULS
DUT
SUL
LIT
CONN.
PUT
FAI
PIK
ORA
ROC
WES
MON
SUS
CAR
WAR
MOR
HU
NAS
SUF
NOR
HUN
NY.
LEH
PA.
BUC
MER
MON
NJ.
MOT
CHE
PHI
OCE
BUR
BEL
GLO
CAM
NC
SAL
ATL
CUM
CAP
DEL.
*
HU
GROUPS
▦ Urban/Industrial Core
▩ Polish/Italian Corridor
▥ Low Socioeconomic States
☐ One or more low risk groups for the male population
☐ No Distinct Group

Thousands of bivariate correlations were computed in the course of screening the high risk indicators. The most important correlations are between the 24 cancer categories and the 62 final high risk indicators. The bivariate correlations between these two matrices are presented in Table 31 at the start of this chapter for three geographical regions: (1) entire study area 45 political entities; (2) 40 places (New York City, Philadelphia, Carbon, Pike and Monroe counties excluded); and (3) 21 county state of New Jersey. The intercorrelation matrix contains 4140 bivariate correlations. Using an 0.05 confidence level as a cutoff, 1668 significant correlations (40 percent) were found. About 10 percent of these significant correlations were considered to be unexpected. Only 207 could have been expected to occur by chance. Specifically, 156 correlations were classified as potentially false positives, based on the literature and the expected associations stated in the first chapter.

It is important to note, however, that our initial categorization of a correlation as unexpected does not mean that a real association does not exist. For example, when viewed from from the perspective of the 24 cancer categories, almost eighty percent of the potentially false positive results are with female white breast cancer (Table 42).

TABLE 42

Pattern of Correlations Classified as Unexpected

Body Site	Number	High Risk Indicator	Number
White female breast	120	Personal	35
Skin	12	Occupational	95
Other	24	Local Ambient	26
Total	156	Total	156

The female breast cancer associations with environmental and ethnic factors are so consistent and strong that it is hard to accept the associations as spurious. Some of the other positives also may not be spurious. For example, we expected socioeconomic status to be inversely correlated with stomach, liver, and bladder cancer. These expected correlations appeared, as did inverse associations with cancer of the rectum and pancreas. The rectum and pancreas correlations were classified as unexpected because we found no literature to support negative association hypotheses. But they may be real associations. Chapter 4 will review selected correlations initially classified as unexpected by body site.

Major differences were found in the number of significant correlations with high risk indicators. All cancer sites, stomach, and rectum each had more than 250 significant correlations. Cancer of the prostate, bladder, and skin each had less than 25 significant associations.

Differences in the tendency to associate by body site were paralleled by differences in the importance of the independent variables. Male median education completed, the proportion of males completing at least college, and the proportion of workers employed in transportation equipment manufacturing each demonstated less than six significant correlations. Thirty-three high risk indicators exhibited more than thirty significant correlations. Twenty-three of the thirty-three are occupational variables and seven are air environment indicators. The remaining three are percent Polish, percent Italian and population density. The TSP/wind direction/distance index demonstrated the greatest number of significant correlations, 37.

Summary

Summarizing, this chapter has described how 62 indicators of high cancer risk were passed through mathematical screening devices to form the basis of multivariate analyses. In the following chapter, we turn to a multivariate analysis of the association beetween cancer mortality and high risk indicators.

CHAPTER 4

RESULTS BY DISEASE

This chapter presents the statistical associations between white cancer mortality and high and low risk factors. The analysis is restricted to white cancer mortality because as previously indicated the nonwhite data set did not have enough cases to warrant a detailed analysis. Furthermore, nonwhites are not a homogeneous population. They include Asians and other races as well as blacks. For these two reasons, a nonwhite study was judged to be infeasible.

The vast majority of the chapter consists of a review of 24 sex/age/and race standardized cancer categories and the three cancer factors presented in Chapter 2. The 24 standardized cancer categories are presented, then the three factors. The order of presentation of the 24 sites reflects the number of significant bivariate correlations found for the 45 case (entire region), 40 county (excluding New York City, Philadelphia, Carbon, Monroe, and Pike), and 21 county (New Jersey) regions. The 35 and 25 county correlations are presented only when they are very different from the 45, 40, and 21 area correlations. Those categories with the fewest significant correlations are presented first. The 24 sites are presented in reverse order of the number of significant correlations because the greated the number of significant correlations, the more complex the interaction between the high risk factors. It was felt desirable to present the more complicated results later in the chapter, and the less complicated first.

The maximum number of significant correlations is 171 for the white male population (57 indicators x 3 regions) and 174 for the white female population (58 indicators x 3 regions) (Chapter 3-Table 31). The number of significant correlations by body site varies from as few as 2 to as many as 140. Nine correlations significant at the .05 level could be expected to occur by chance; 2 might be expected at the .01 level. All of the correlations significant at the .05 level are in Table 31.

Whereas the diseases are introduced in order of the least number of significant correlations, the body sites are not separated. Since in many cases the same high risk factors are relevant to the male and female populations, we did not separate the body sites. For example, male large intestine exhibited many significant correlations. It is discussed midway into the chapter, rather than near the end, because female large intestine cancer mortality had far fewer correlations.

The chapter provides similar information about each body site. Initially, the relative importance of the disease is summarized. Then, expected associations between the body site and high risk factors are reviewed. Third, the results are discussed. The bivariate correlation results are presented, then the multivariate analyses. Efforts are made to indicate cases when one independent variable is probably acting as a surrogate for others. Finally, future research efforts suggested by the results are reviewed.

In order to be consistent, stepwise multiple regression results are always presented. The stepwise method chooses the most significant indicator first. Then subsequent indicators are selected as long as the multiple correlations coefficient increases and the correlation between the high risk factors is not so high as to preclude separate interpretation of the indicators. We controlled the number of independent variables that may enter by requiring each variable to pass a probability of significance test.

In this study many of the high risk indicators are correlated. Therefore, few multivariate analyses produced more than two or three high risk indicators. Because the high risk indicators chosen by the mathematical method may represent many high risk factors, we have tried to point out the important possibility of oversimplification and misinterpretation.

While the stepwise results are presented to ensure consistency, other multivariate procedures were also used. For example, we sometimes used an option that gave us the best single indicator, best pair of high risk indicators, and best three indicators. We also sometimes specified the order of variables to determine if indicators of special interest to us would still be incorporated into the results. We chose to present only the stepwise results to spare the reader interpretative indigestion. Overall, we have tried to be extremely careful in order to avoid misleading results. Nevertheless, the results of the analyses presented in this chapter cannot be taken at face value. Many of the high risk indicators are approximations and potentially important high risk factors such as smoking and alcohol consumption are not incorporated into the analysis due to a lack of data.

White Male Prostate Cancer Mortality

Prostate cancer is responsible for more than nine percent of male cancer mortalities in the region. Despite its importance, we used prostate as a control disease. It was not expected to and did not strongly correlate with any of the high risk factors.

Only 2 of 171 correlations were significant (Table 43) at the .05 level.

TABLE 43

Significant Bivariate Correlations:
White Male Prostate Cancer Mortality, 1950-1969

Variable	40 Areas R	21 Areas R
1. % same house	.264	-
2. % chemical workers	-	.381

Are chemical workers (exposed to cadmium?) subject to higher rates of prostate cancer in New Jersey? Perhaps they are; our evidence, however, does not warrant such a conclusion. Following the guidelines outlined in Chapter 1, we conclude that the two statistically significant correlations are spurious. If research on prostate cancer is to be conducted, the major difference between white and nonwhite rates would seem to be the logical starting place.

White Female and Male Bladder Cancer Mortality

Bladder cancer was responsible for less than two percent of female cancer mortalities in the region during the 1950-1969 study period. We expected inverse correlations with socioeconomic status and the presence of the four ethnic groups. Air and water pollution were other possible high risk factors.

Only 2 of 174 correlations were significant (Table 44), and these correlations appeared only for the 40 case study area.

TABLE 44

Correlation Analysis,
White Female Bladder Cancer Mortality, 1950-1969

40 Areas

Variable	R	Mult R	R^2	Change R^2
1. % German population	-.319	.319	.102	.102
2. % Chemical workers	.299	.389	.151	.049

The German association was expected, the chemical association was not. Socioeconomic status correlated in the expected direction, but not significantly.

The stepwise program accepted both the percent German and percent chemical indicators as significant. Together they account for 15 percent of the variance (Table 44).

While we cannot report much success in explaining the spatial pattern of white female bladder cancer deaths in the region, the spatial pattern is noteworthy. The study area's rate exceeds the national rate by 18 percent. Four of the five and eight of the ten highest rates in the study area are found in New Jersey. The four highest New Jersey rates area in southern New Jersey: Gloucester, Cape May, Salem, and Camden.

Four hypotheses have come to mind to explain the geographical concentration. Perhaps the southern New Jersey chemical industry, centered in Salem County, is involved beyond the association with occupation. When Salem County is dropped from the analysis, the variable percent chemical workers drops out of the association. Therefore, we can conclude that the chemical industry, if it is involved, is probably limited in association to Salem County, and possibly some commuters to Gloucester and Camden Counties. Even this conclusion is probably too strong because we have no information about female participation in the chemical industries in these counties; nor do we possess information about the capability of chemicals used in these plants to be spread from men's working clothes to women at home; and we have no data implying that pollutants are diffused to the north, east and northeast in the air or into water supplies. In short, the chemical hypothesis is possible, but tenuous.

The second hypothesis is that one or more of the missing variables, for example, smoking, are the keys. One could hypothesize that people in southern New Jersey smoke more than people elsewhere in the region. The smoking hypothesis is, however, rejected if female lung cancer is used as a surrogate for smoking. The correlation between white female bladder and lung cancer in the region is an insignificant -.081. Furthermore, our current efforts to develop smoking information at the county scale suggest that less smoking should be expected in the southern New Jersey counties than in the region as a whole.

The third possibility is that an unrecognized local ambient environmental factor (such as pesticides in the water supply) is involved. Our percent agricultural worker variable was not significantly correlated with white female bladder cancer mortality. The indicator may be too general. However, it is all we have at the present time.

A fourth hypothesis is that through intermarriage a genetic tendency has been passed along in these counties. The hypothesis is interesting, but we have no evidence to support or reject it. Overall, we have no evidence to support or to reject the smoking, intermarriage, the water supply or chemical industry hypotheses.

Good possibilities of testing them exist. Specifically, Cumberland County which lies adjacent to Salem and close to Gloucester, Cape May, and Camden Counties, had one of the lowest white female bladder cancer rates in the region (2.3/100,000 compared to between 3.4 and 3.7 for Salem, Gloucester, Cape May, and Camden). Similarly, Warren County (3.2/100,0000) and adjacent Sussex County (1.4/100,000) have different rates. Before strongly suggesting a special study for a relatively unimportant cause of female cancer mortality, we suggest examining bladder cancer rates for the 1970's in order to determine if the southern New Jersey female bladder cancer concentration has remained high. Indeed, 1968-72 data which will be presented in Part II show a substantial change in southern New Jersey.

The study area has a high white male bladder cancer mortality rate (9.24/100,000). The regional rate exceeds the upper decile of the states. The rates vary from 16.1/100,000 in Salem County, New Jersey (reportedly the highest in the United States) to 6.5/100,000 in Chester County, Pennsylvania.

White male bladder cancer rates significantly correlated with only four of the other 23 age/race/sex standardized cancer mortality categories: female bladder, male leukemia, and male skin. Since exposure to specific chemicals is thought to be one of the causes of leukemia and skin cancer, an occupational exposure hypothesis seemed reasonable. Overall, we expected to find inverse associations with socioeconomic status and the presence of Italian populations and positive correlations with the presence of industry and agriculture, especially the chemical industry, and air and water contamination.

As in the female bladder case, with the exception of Pike County, Pennsylvania, the highest bladder rates are found in New Jersey (Salem, Cape May, Ocean, and Hunterdon). Two of the four high New Jersey counties are in the set of high female bladder cancer counties (Salem, Cape May). The lowest bladder cancer counties are outside of New Jersey (Carbon , New Castle, Dutchess, Montgomery, Chester). Only Dutchess County is common to the low female and low male bladder cencer mortality counties. Smoking may be associated with bladder cancer. However, as in the case of the female white population, the correlation between male bladder and lung cancer is insignificant. Finally, it is interesting to note that none of the high or low rate male or female bladder cancer counties is heavily urbanized.

The statistical analysis of white male bladder cancer mortality is among the most complicated. Nevertheless, the author believes it is important to follow it through some of the steps. Thirteen of the 171 correlations were significant at the .05 level (Table 45). In addition, the correlations of the 25 most populous counties are presented because they add another dimension to the presentation.

TABLE 45

Significant Bivariate Correlations: White Male Bladder Cancer Mortality, 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R	25 Areas R
1. % chemical workers	.409*	.456*	.541*	-
2. chemical workers/area index	-	-	-	.528
3. % professional workers	-.353*	-.401*	-	-.359
4. % male at least completed college	-.339	-.392*	-	-
5. Male median education	-.287	-.333	-	-
6. Median income	-	-.268	-	-
7. % agriculture workers	-	-	.398	-.396
8. % Italian	-	-	-.370	.377
9. Socioeconomic status factor	-.328	a	a	a
10. % Polish	-	-	-	.366
11. % U.S.S.R.	-	-	-	.373
12. All manufacturing workers/area index	-	-	-	.385

* Significant at .01 level

a Only considered for n=45 case

Chemical workers is the only variable that consistently correlates across the three geographical regions. Professional workers is significant in three of the four regions and almost significant in the fourth. We regard the professional variable as a surrogate for high socioeconomic status. The 45 and 40 case study areas manifest similar correlations with chemical workers and indicators of socioeconomic status. In New Jersey the affluence indicators drop out and are replaced by agricultural worker and the expected negative association with Italian populations. All of the statistically significant indicators were expected; but they were expected to be consistent and are not.

The geographical inconsistency required separate analyses of the region and subregions. With respect to the 45 and 40 case region and subregion, the chemical and professional variables account for more than one third of the variation in the male white bladder cancer mortality variable (Table 46).

TABLE 46

Multivariate Analysis of White Male Bladder Cancer Mortality, 1950-1969, 45, 44, and 40 Area Regions

	45 Areas			40 Areas			(Without Salem County) 44 Areas		
Variable	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2
1. % chemical workers	.409	.167	.167	.456	.208	.208	-	-	-
2. % professional workers	.585	.342	.175	.639	.408	.200	.353	.124	.124

The correlation between chemical workers and professional workers is an insignificant .149. Accordingly, there is little multicollinearity between the two independent variables. Overall, high risk counties for males in the study area seem to us at first to have a large proportion of chemical workers and a relatively low socioeconomic status population.

Upon further analysis, we found that if Salem county is dropped from the region, the chemical workers variable drops from the relationship and only the professional variable is retained (Table 46, 44 areas). In short, the chemical worker - male white bladder cancer link at first seemed to be tied to Salem County, New Jersey.

When the region is New Jersey's 21 counties, the same conclusion is drawn. Initially the chemical, professional and Italian variables were found to be important (Table 47). The agricultural variable was correlated with the chemical variables and therefore was rejected. The professional variable was not significant at the .05 level. However, it was inserted into some of the preliminary multivariate tests and proved to be important.

Table 47

Multivariate Analysis of White Male Bladder Cancer Mortality, 1950-1969, 21 and 20 Area Regions

	21 Areas			(Without Salem County) 20 Areas		
Variable	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2
1. % chemical workers	.541	.293	.293	-	-	-
2. % professional workers	.660	.435	.143	.289	.084	.084
3. % Italian population	.713	.508	.072	.403	.163	.079

When Salem County is included, the proportion chemical workers indicator is *highly significant* and consistent. The stepwise program chose it first and it accounted for 29.3 percent of the variance. When we forced percent chemicals in last, it accounted for 28.9 percent of the variance.

However, when Salem County is deleted, the chemical variable drops out. Indeed, with the exception of Salem County, the other counties that have chemical industry oriented labor forces do not have among the highest rates of white male bladder cancer in New Jersey (Table 48). Somerset and Middlesex Counties have relatively low rates, while Union County has a higher rate. Do or did Salem and Union Counties have in common the production of some chemicals (dyes?) not produced in Middlesex and Somerset Counties?

TABLE 48

Comparison of Proportion of Chemical Workers and Male White Bladder Cancer

County	Avg. % chemical workers 1950-1970	Rank in New Jersey	Male White Bladder/100,000 Cancer Mortality 1950-1969	Rank in New Jersey
Salem	26.5	1	16.1	1
Somerset	10.5	2	7.9	20.5*
Middlesex	10.3	3	8.4	18
Union	7.1	4	9.9	9
Gloucester	7.0	5	9.2	12**

*Tied with Warren County for lowest

**Tied with Mercer and Burlington Counties

Yet when the correlations were made on the 25 most populated counties (accounting for the overwhelming majority of the region's population) the chemical factor again returned. The chemical workers/area index produces the highest correlation (.528), a correlation significantly higher than any other. In addition, the sign of the Italian and agricultural workers variables change. In short, the correlation analyses are quite inconsistent.

The residuals from regression analyses strongly suggested that the high rates in Cape May and Pike Counties were independent of the high risk indicators used in the multivariate analyses. These two counties had much higher observed than predicted bladder cancer rates. As indicated in Chapter 2, we believe that intra-regional migration is responsible for the high cancer rates in these two counties. Doctors in the two counties were contacted in an effort to gather some evidence. Two Pike County doctors remembered that two of the three last male bladder cancer victims had previously been residents of metropolitan New York and Philadelphia. However, the county medical officials were unable to trace the birthplace, previously living location, and occupations of these individuals. Officials of Shore Memorial Hospital, Somers Point, New Jersey agreed to provide a small sample of bladder and large intestine cancer cases. Among the six male bladder cancer victims, five were residents of Atlantic County and only one a resident of Cape May County. The one Cape May resident was born in Philadelphia and had worked for a refinery for 40 years. The single case supports our hypothesis, but only suggests that an epidemiological study of cancer mortality in Cape May and probably Pike Counties would be interesting.

Summarizing, four expected high risk factors were found to be statistically associated with white male bladder cancer mortality. Low socioeconomic status is moderately correlated with high bladder cancer rates. The proportion of chemical workers among a county's labor force is strongly associated with white male bladder cancer. When Salem County is removed, the correlation becomes insignificant. But when only the 25 most populated counties are considered, the significant correlation reappears. The presence of a large proportion of Italian nationals lowers the male white bladder cancer mortality rates in one case and seems to increase in another. Finally, we hypothesize but cannot demonstrate that the high rates in Pike and Cape May Counties are due to intraregional migration.

The association between white male and female bladder cancer mortality is relatively weak: R is .483 in the entire region and .185 in New Jersey. Salem, Cape May, and Pike Counties had among the five highest male and female rates. We have no firm statistical evidence linking the relatively high male and female rates in these counties.

In conclusion, this study provides some support for and against the commonly heard statement that the region's high white male bladder cancer rate is linked to chemical industrial exposure. Perhaps we have not included the appropriate industries, or the industrial information was too general. Alternatively, the region's high male and female bladder cancer rates may not be due to industrial exposure.

While we have suggested further research on bladder cancer, it seems paramount to determine from more recent cancer rates if Salem County has continued to manifest high rates and if other chemical oriented counties have changed their bladder cancer rates. The 1968-1972 data to be presented in Part II will further complicate these preliminary findings.

White Female and Male Skin Cancer Mortality

Many people contract skin cancer. Fortunately, few people die from skin cancer. The region's skin cancer mortality rate is 95 percent of the national average (1.72 white females and 2.91 white males).

We expected positive correlations between white female skin cancer and outdoor occupations (agriculture) and negative correlations with socioeconomic status. Socioeconomic status serves as a general indicator of indoor occupations. Agriculture is a surrogate for outdoor exposure and exposure to arsenic and hydrocarbon based pesticides. The same associations were expected with white male skin cancer deaths. In addition, we expected to find a significant positive correlation with chemical workers and construction workers.

As expected, the highest female skin cancer rates are found in rural counties (Sussex, Cape May, Warren, Hunterdon, and Sullivan). The lowest rates, however, also are largely rural and exurban counties (Monroe, Rockland, Bucks, Litchfield, Dutchess, Northampton, and Pike). The 8 bivariate correlations suggest extremely weak verification of the outdoor and income hypotheses and the unexpected addition of ethnic factors (Table 49).

TABLE 49

Significant Bivariate Correlations:
Female White Skin Cancer Mortality, 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R
1. Median Income	-	-.296	-
2. % professional workers	-	-.282	-
3. % Italians	-	-.256	-.480
4. % Agricultural workers	-	.252	.520*
5. % Polish	-	-	-.435
6. % USSR	-	-	-.401

*Significant at .01 level

Only 8 of 174 correlations were significant. Nine would have been expected by chance. None were significant in the 45 case study area. The stepwise multivariate program accepted the agricultural worker and median income variables. In the 40 county region, these two indicators accounted for less than 12 percent of the variation in white female skin cancer mortality (Table 50). The only significant correlation in the New Jersey region was agricultural workers (Table 50).

TABLE 50

Multivariate Analysis of White Female
Skin Cancer Mortality, 1950-1969

	40 Areas			21 Areas		
Variable	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2
1. Median Income	.296	.087	.087	-	-	-
2. % Agricultural workers	.346	.120	.033	.520	.270	.270

Do the three ethnic groups implicated in the bivariate correlation studies have lower skin cancer rates because of their pigments? Or, as the multivariate analysis suggests, is agriculture a confounding factor? We have assumed, perhaps incorrectly, that the negative ethnic correlations are spurious and result from the fact that these foreign born groups usually live in urban areas and are not engaged in agricultural occupations. The correlations between the three nationality groups (Polish, Italian, and USSR) and percent agricultural employment range from - .488 to - .693. Overall, the high risk white female skin cancer mortality counties, especially in New Jersey, contain concentrations of agricultural employment.

Agricultural counties also demonstrate higher rates of white male skin cancer mortality. However, the presence of chemical industry completely overshadows every other high risk indicator. Agricultural and construction workers are insignificantly correlated at all three geographical scales. We did not collect fishermen data, another possible occupational exposure group. However, the results suggest that fishing exposure would not be a significant factor.

Only 9 of 171 correlations are significant. They depict counties with a chemical economic base and secondarily stable populations containing few German nationals (Table 51).

TABLE 51

Significant Bivariate Correlations:
White Male Skin Cancer Mortality, 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R
1. % chemical workers	.573*	.616*	.702*
2. % German population	-.280	-.298	-.494
3. % same house	-	.310	.398
4. % managerial workers	-	-	-.369
5. % agricultural workers	.164**	.165**	.269**

*Significant at .01 level
**Not signfificant at .05 level

The stepwise multivariate method selected the chemical workers and the German population variables in the 45 and 21 case analyses (Table 52). The chemicals variable is responsible for 33 and 49 percent of the variation in the 45 and 21 case study areas. The German and stability variables add slightly to the explanatory power of the results. The German population correlation was not expected and is therefore tentatively classified as spurious. However, readers of this volume may have information to support a hypothesis that German foreign stock Americans are less prone to skin cancers than other Americans.

TABLE 52

Multivariate Analysis of White Male Skin Cancer Mortality, 1950-1969

				(Without Salem County)					
	45 Areas			40 Areas			21 Areas		
Variable	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2
1. % chemical workers	.573	.329	.329	.309	.096	.096	.702	.493	.493
2. % German population	.595	.354	.025	.385	.128	.032	.773	.598	.105
3. % same house	-	-	-	-	-	-	.794	.631	.033

Next, we determined if the positive correlation with chemical workers was solely dependent on Salem County. It was not. Upon deleting Salem County chemicals was still chosen first by the stepwise procedure. While its explanatory power dropped from over 30 percent to about 10 percent, the relationship is not totally dependent upon Salem County. Middlesex and Somerset Counties have high white male skin cancer mortality rates and have averaged more than 10 percent of the labor force in chemical production.

A suggested project based on the multivariate analysis results is to determine if in fact the mortalities in Salem, Middlesex and Somerset were concentrated disproportionately among chemical workers. If the answer is yes, then the next step is to determine if chemicals carcinogenic to the skin (arsenic, petroleum, coal products, chlorinated hydrocarbons?) have been produced in these counties, while other counties that have had economic bases relying heavily on the chemical industry (Union and Gloucester) are used as controls.

Summarizing, skin cancer mortality has not been a major cause of death in the region. High risk counties for females have a slight agricultural orientation and relatively low to lower middle income populations, while high risk male counties tend to be relatively specialized in chemical production.

White Female and Male Leukemia Cancer Mortality

The region's leukemia rates are slightly below the national average. Leukemia was responsible for about 4.3 percent of cancer mortalities among white females and males in the region during the 1950-1969 study period. While it is not a major cause of mortality, the disease is interesting to analyze because it gave us the opportunity to evaluate a strong ethnic hypothesis. Specifically, we expected to find a strong association with our Jewish surrogates (% Russian, % German).

Three other hypotheses were put forth. One is that affluent people are subjected to more x-rays and therefore have a higher leukemia risk. The second is that males in the petrochemical industry are exposed to the carcinogen benzene. A related, abet more remote, possibility is that air pollution has distributed the ubiquitous chemical throughout urbanized areas. Finally, the strong possibility existed that the leukemia cancer rate was not an accurate indicator of the spatial distribution of environmentally caused leukemia because genetics may play an especially important role in pre-adult cases.

The results weakly support the Jewish hypothesis. With respect to the white female population only 5 of 174 correlations were significant (Table 53). But the 25 case (most populated counties) produced 30 significant correlations of which the USSR indicator is by far the strongest correlate. The vehicle mile/wind direction/distance index and air emissions/wind direction/distance index also exhibited high correlations.

TABLE 53

Significant Bivariate and Multivariate Correlations:
White Female Leukemia Cancer Mortality, 1950-1969

Variable	45 Areas	40 Areas				25 Areas
	R	R	Mult R	R^2	Change R^2	R
1. % different house, same county	.345*	-	-	-	-	-
2. % managerial workers	-	.300	0.300	.090	0.090	-
3. % German population	-	.271	-	-	-	-
4. % USSR population	-	.269	-	-	-	.610
5. Female median education	-	.255	-	-	-	-
6. VMT/wind direction/ distance index	-	-	-	-	-	.576
7. % Italian	-	-	-	-	-	.338

*Signficant at .01 level

The 45 area region produced a significant correlation only with a stability variable. The stepwise program for the 40 area region accepted only the highest correlation, managerial workers, because the independent variables are correlated.

The bivariate correlations between percentage managerial workers and proportion Germans, Russians, and female median education are .498, .471, and .560. By comparison, the correlations between managerial workers and the Polish and Italian variables are insignificant. Overall, we feel that the correlation analysis supports the Jewish/leukemia link, albeit weakly.

Before leaving female leukemia mortality, we note an interesting unexplained geographical pattern. The managerial/Jewish based regression equations' most glaring underpredictions were in New Jersey: Hunterdon, Sussex, Salem, Essex, and Mercer Counties. In addition, Warren County's female leukemia rate was underpredicted. Is there a reason why New Jersey, in general, and the four northwestern New Jersey counties, in particular, had higher rates than predicted by the model?

The higher than expected leukemia rates in northwestern New Jersey are half of the puzzle. With the exception of Suffolk County, New York, the five counties in which the rates were lower than expected are adjacent to one another in southern New Jersey: Cumberland, Cape May, Gloucester, and Burlington. Could there be a common cause(s) in southern New Jersey? Perhaps, once having removed the Jewish population as a high risk group, there remains a systematic factor(s) that cause northwestern New Jersey counties and southern New Jersey counties to experience the greatest differences in female leukemia cancer mortality in the region. Lynne Harrison (note 1978) suggests that the finding of clusters is one of the reasons why some epidemiologists believe that leukemia has an infectious, virus basis.

The association of male leukemia and risk factors is stronger and more complex than the female population, but strongly tied to the Jewish population. Forty-two correlations were significant. Most were eliminated by applying the consistency and fallback rules (Table 54).

TABLE 54

Important Bivariate Correlations:
White Male Leukemia Cancer Mortality, 1950-1969

	Variable	45 Areas R	40 Areas R	21 Areas R	25 Areas R
	No. of Significant Correlations	7	15	20	19
1.	% USSR population	.477*	.540*	-	.504
2.	VMT/wind direction/ distance index	.340	.454*	-	.501
3.	% managerial workers	.309	.279	-	.389
4.	% German population	.256	-	-	-
5.	Construction workers	-	.410*	.486	-
6.	Manufacturing workers	-	.330	.485	-
7.	% Italian	-	-	-	.545
8.	Control Variable: population density	-	.307	.440	-
9.	Urban/Industrial core factor	.318	a	a	a

* Significant at .01 level
a Only considered for n=45 case region

Seven significant bivariate correlations were found for the 45 case study area, 15 for the 40 case region. The expected correlation with the relative size of the USSR ethnic population was clearly dominant. The stepwise multivariate program accepted the USSR variable (Table 55). The managerial indicator was accepted in the 45 case region and rejected in the 40 case study area. High corrrelations among the independent variables prevented any of the other variables from entering. Overall, the USSR variable explained about 25 percent of the variance of white male leukemia cancer mortality (Table 55).

TABLE 55

Multivariate Analysis of White Male
Leukemia Cancer Mortality, 1950-1969

	45 Areas			40 Areas			21 Areas		
Variable	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2
1. % USSR population	.477	.227	.227	.540	.292	.292	-	-	-
2. % managerial workers	.502	.252	.025	-	-	-	-	-	-
3. Manufacturing workers	-	-	-	-	-	-	.485	.235	.235

The Jewish (U.S.S.R.) factor did not appear in New Jersey. Twenty significant correlations focusing on urbanization and industrialization were found. These 20 significant bivariates include positive associations with the location of the construction, chemical, fabricated metal, machinery, and transportation equipment industries. Using all manufacturing employment and population density as control variables, only the construction industry indicator was retained. We have no reason to expect a particularly high correlation with construction workers, as opposed to other workers, and the construction industry correlation is virtually identical to the all manufacturing indicator. The correlation between the two is .898. Therefore, the all manufacturing variable was retained and used as the urban/industrial surrogate.

Summarizing, we expected and found significant correlations between leukemia and surrogates for the presence of a Jewish population. Affluence and the presence of the chemical industry were expected to be significant. They were significant. Finally, the vehicular traffic and Italian indicators were significant and strong in some cases and insignificant in others. There is no single outstanding county for further leukemia research. New York City, Philadelphia, Essex, Westchester and Nassau Counties all have high rates. Indeed, before further research is undertaken, we suggest recalculating and studying the leukemia rates for the age group 30-75. This is being done for the Part II study.

White Female and Male Kidney Cancer Mortality

The region's female and male kidney cancer mortality rates exceed the national averages by 4 and 10 percent, respectively. Mortality due to kidney cancer is a relatively minor cause of mortality (2.07/100,000 for females and 4.24/100,000 for males in the region). We expected low positive associations with ethnicity (Polish, USSR), and perhaps degraded air and surface water quality.

The results matched the modest expectations. With respect to the female population, 5 of 174 correlations were significant (Table 56). The surface water and Polish variables are correlated (R=.437). Therefore, the stepwise program accepted only the surface water variable. While neither indicator is strongly correlated with kidney cancer, sorting out the interaction between the two high risk factors is important from the policy perspective. On the one hand, if elevated kidney cancer rates are due to contaminated water supplies, then government can choose to act. On the other hand, if the elevated rates are due to unique ethnic habits, then government actions are restricted.

TABLE 56

Bivariate and Multivariate Analyses of White Female
Kidney Cancer Mortality, 1950-1969

Variable	45 Areas R	45 Areas Mult. R	45 Areas R^2	40 Areas R	40 Areas Mult. R	40 Areas R^2	25 Areas R
1. % potable surface water supply, 1963	.318	.318	.101	.335	.335	.112	.447
2. % Polish population	.254	-	-	.284	-	-	.510
3. % self service water supply, 1963	-	-	-	-.267	-	-	-.631

Judging from the results of the white male kidney analysis, the water supply relationship is tenuous, the ethnic association a stronger possibility. Specifically, among the 59 significant correlations with white male kidney cancer, the Polish and stability variables are the most consistent (Table 57).

TABLE 57

Important Bivariate Correlations:
White Male Kidney Cancer Mortality: 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R
No. of Significant Correlations	3	22	34
1. % Polish population	.306	.438*	.522*
2. % USSR population	.249	.392*	-
3. % same house	.383*	.357	.426
4. manufacturing workers	-	.300	.551*
5. construction workers	-	.312	.568*
6. chemical workers	-	.360	.594*
7. fabricated metal workers	-	-	.618*
8. VMT	-	-	.680*
9. Median income	-	.280	.772*
10. % potable surface water, supply 1963	-	-	.405
11. % self service water supply, 1963	-	-	-.524*
12. % transportation equipment manufacturing workers	-	-	.535*
13. TSP/wind direction/distance index	-	.344	.445

*Significant at .01 level

The USSR indicator is the third strongest in the 45 and second strongest in 40 case regions. Only in the 21 county New Jersey area are other high risk factors more significant. Income and education (socioeconomic status) variables produced higher positive correlations than the Polish population. In addition, industrial and local ambient environmental indicators appear stronger in New Jersey than the Polish indicator. Are the occupational and environmental variables contributing factors? Or do they all contribute? Based on the hypotheses formulated in Chapter 1 and consistency guidelines, we have to conclude that the ethnic factors contribute. The socioeconmic status, occupational, and local ambient environmental factors may increase the risk or may be spurious.

Examination of the highest and lowest rates did not help us to explain further the intraregional variance in white male kidney cancer mortality. Among the counties with the highest rates three are rural (Pike, Monroe, Sullivan) and three are more urbanized (Union, Nassau, and Camden). The counties with the lowest rates are relatively rural (Cape May, Ocean, Dutchess, and Chester).

The multivariate analysis for the 45 case region chose the same house variable as primary (Table 58). The income indicator accounted for most of the variance in the New Jersey study area. Intercorrelation among the independent variables is strong. Accordingly, the VMT and percent transportation manufacturing variable only play a minor role.

TABLE 58

Multivariate Analyses of White Female Kidney Cancer Mortality, 1950-1969

Variable	45 Areas			40 Areas			21 Areas		
	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2
1. % same house	.383	.147	.147	.526	.277	.085	-	-	-
2. % Polish population	.471	.222	.075	.438[a]	.192[a]	.192[a]	(.522)**	(.272)**	(.272)*
3. % USSR population	.494	.244	.022	.584	.341	.084	-	-	-
4. Median income	-	-	-	-	-	-	.772	.596	.596
5. VMT	-	-	-	-	-	-	.825	.681	.085
6. % employed in transportation equipment manufacturing	-	-	-	-	-	-	.862	.743	.062

** Not incorporated in the multivariate model

[a] % Polish should be read first, % same house read second

Overall, the ethnic factor is the most persistent indicator. Other personal, occupational, and local ambient factors are stronger only in New Jersey. The strength shown by these non-ethnic factors may be spurious.

There is no single or group of counties that stand out as the place to study kidney cancer mortality. The most interesting case is Monroe County. Monroe has one of the highest white male and lowest white female rates.

White Female and Male Pancreas Cancer Mortality

The study area's female and male pancreas cancer mortality rates are respectively 7 and 4 percent higher than the national averages. As in the case of prostate and leukemia cancer mortality rates, we had doubts about the accuracy of the mortality rates. Smoking may be another confounding factor since pancreas cancer has been linked with smoking. The correlations between pancreas and lung cancer,however, are insignificant. Assuming the rates are accurate, we expected weak associations with USSR and Polish populations and for the male population an association with the presence of the chemical industry.

The significant correlations with white female pancreas cancer mortality present an unclear picture. The Polish and USSR correlations are significant only when the 25 most populated counties are the study region. They were between 0.0 and 0.24 in the other regional groupings. Other correlated factors are socioeconomic status, stability, and air quality (Tables 59 and 60).

TABLE 59

Important Bivariate Correlations of White Female Pancreas Cancer Mortality, 1950-1969

Variable No. of Significant Correlations	45 Areas R 48	40 Areas R 6	21 Areas R 4	25 Areas R 28
1. VMT/wind/direction/distance index	.371*	.310	-	.658
2. % professional workers	-.323	-.409*	.376	-.425
3. % different house, same county	.290	.264	.458	-
4. % same house	.257	.438*	.500	.389
5. % female completed at least colege	-.254	-.358	-.399	-
6. Urban/industrial core factor	.265	a	a	a
7. % Polish	-	-	-	.506
8. % USSR	-	-	-	.526

* Significant at .01 level
a Only considered for n=45 case

TABLE 60

Multivariate Analysis of White Female Pancreas Cancer Mortality, 1950-1969

Variable	45 Areas Mult. R	R^2	Change R^2	21 Areas Mult. R	R^2	Change R^2
1. VMT/wind direction/distance index	.371	.138	.138	-	-	-
2. % professional workers	.492	.242	.104	-	-	-
3. % same house	-	-	-	.500	.250	.250
4. % female completed at least college	-	-	-	.618	.382	.132

Studying the high rate counties does not help. The high rate counties include urban New York City and Nassau and rural Warren, Salem, and Litchfield. Overall, if there are important clues to be found in the statistical analysis of white female pancreas cancer, they certainly are inconsistent.

In contrast to the confusing analysis of white female pancreas mortality rates, the white male results generally followed the expected pattern. The strongest of the 41 significant associations are with ethnic groups (Table 61). The chemical workers variable is significant, but inconsistent, and not as strong as the ethnic and air pollution factors. However, it is stronger than any of the other manufacturing categories.

TABLE 61

Important Bivariate Correlations: White Male Pancreas Cancer Mortality: 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R	25 Areas R
No. of Significant Correlations	23	12	12	7
1. % USSR population	.456*	.462*	-	.622
2. % Polish population	.412	.454*	-	.573
3. VMT/Wind direction/distance index	.360*	.368*	-	.440
4. Chemical workers	.320	.283	-	-
5. Manufacturing workers/ distance index	.285	-	-	-
6. Population density	.265	-	-	-
7. % Italian population	-	-	.439	.403
8. Urban/industrial core factor	.324	a	a	a
9. % German population	-	-	-	.330

* Significant at .01 level
a Only considered for n=45 case

The stepwise multivariate analysis selected the ethnic indicators (Table 62). While the USSR and Polish correlations were expected, the Italian was not. We suspect but are not convinced that it is spurious. In addition, we suspect that intraregional migration may be an intervening factor because Cape May and Pike have two of the highest rates.

TABLE 62

Multivariate Analysis of White Male Pancreas Cancer Mortality, 1950-1969

Variable	45 Areas Mult. R	R^2	Change R^2	21 Areas Mult. R	R^2	Change R^2
1. USSR population	.456	.208	.208	-	-	-
2. % Polish population	.543	.295	.087	-	-	-
3. % Italian population	-	-	-	.439	.193	.193
4. Urban/industrial core factor	.324	.105	.105	a	a	a

a Only considered for n=45 case

Overall, the most consistent and strongest correlates of white male and female pancreas cancer mortality are ethnic (USSR and Polish) and the VMT/wind direction/distance index.

White Female Breast Cancer Mortality

White female breast cancer mortality is one of the most important and interesting neoplasms to study. Importance results from the fact that the disease is responsible for almost 20 percent of study area's white female cancer mortalities. Second, the region's white female breast cancer mortality rate exceeds the upper decile for all the states. If the importance of the disease was an insufficient stimulant, the bivariate analysis that produced many seemingly unexpected correlations and the spatial distribution of the disease made this a most interesting disease to analyze.

The analysis began with the recognition that the highest rates were to be found in a circle around New York City. Nassau County to the east, Westchester and Putman to the north, and Essex and Union to the west of New York City had the highest breast cancer mortality rates. The lowest rates are found in Pennsylvania and western and southern New Jersey. The rates in Carbon, Chester, Northampton, Cumberland and Sussex Counties are about two-thirds of the rates in the five highest counties.

The literature suggests that affluence and fertility rates should explain the pattern. The fertility rate indicator was eliminated (Chapter 3) because it did not meet the consistency guideline. The socioeconomic status indicators correlated as expected. However, numerous other occupational and local ambient indicators also correlated (Table 63). Overall, 137 of 174 possible correlations were significant. Of the 137 significant correlations, 120 were considered unexpected.

TABLE 63

Important Bivariate Correlations: White Female Breast Cancer Mortality, 1950-1969

		45 Areas	40 Areas	21 Areas	25 Areas
Variable		R	R	R	R
No. of Significant Correlations		50	44	43	39
1.	socioeconomic status factor	.410*	a	a	a
2.	% managerial workers	.452*	.482*	-	.557
3.	median income	.381*	.496*	.599*	-
4.	female median education	.339	.369*	.376	-
5.	% Italian population	.496*	.501*	-	.564
6.	% German population	.402*	.491*	-	.561
7.	% USSR population	.401*	.335	-	.689
8.	% Polish population	.344*	.360	.439	.557
9.	urban/industrial core factor	.512*	a	a	a
10.	VMT/wind direction/distance index	.573*	.678*	.767*	.628
11.	control variable: population density	.404*	.440*	.521	.433

* Significant at .01 level
a Only considered for n=45 case

The expected correlations with socioeconomic status are represented by four variables in Table 21: the socioeconomic status variable produced by factor analysis; % managerial workers; median income; and female median education. The ethnic group correlations, with the exception of German, were expected to be negative. They were all positive (Table 63, variables 5-8). The more than 100 correlations with occupational and local ambient environment indicators are represented by three variables (Table 63, 9-11): the urban/industrial core factor; the VMT/wind direction/distance index, and the control variable, population density. We use the VMT and the other two indices as surrogates for the occupational and local ambient environment factors.

Recognizing that the ethnic, occupational, and local ambient environmental variables were considered unexpected, we performed the multivariate analyses in four stages; (1) socioeconomic status and German ethnic only; (2) socioeconomic status and ethnic; (3) socioeconomic status, ethnic, and occupational and local ambient environment; and (4) factor analysis indicators.

The initial stepwise multivariate run selected the managerial, income, and German variables (Table 64).

TABLE 64

Multivariate Analysis of White Female Breast Cancer Mortality, 1950-1969, Socioeconomic Status and German Ethnic

Variable	45 Areas Mult. R	R^2	Change R^2	21 Areas Mult. R	R^2	Change R^2
1. % managerial workers	.452	.204	.204	-	-	-
2. Median income	.512	.262	.057	.599	.359	.359
3. % German population	.550	.302	.040	-	-	-

The variables are intercorrelated. Accordingly, the income and German variables only increase the explanatory power of the model by 10 percent. In New Jersey, only income is selected. Summarizing, the expected socioeconomic status indicators account for about one-third of the variance in white female breast cancer mortality.

The second step was to assume that the ethnic correlations were not false positives. Perhaps, the ethnic populations in our study area have higher cancer rates than suggested by the literature. Or perhaps, the foreign born and first generations manifest the expected pattern, but later generations of Americans in these counties have abnormally high rates. Frankly, we doubt the second possibility because the foreign born and first generation should comprise the counties' older citizens. When the ethnic variables are allowed to enter the stepwise program, the Italian and USSR variables are chosen. The explained variance increases to 40 percent (Table 65). Whatever the reason, rates of female breast cancer among the region's ethnic populations are worthy of considerable epidemiological study.

Table 65

Multivariate Analysis of White Female Breast Cancer Mortality, 1950-1969, Socioeconomic Status and Ethnic Factors

Variable	45 Areas Mult R	R^2	Change R^2
1. % Italian population	.496	.246	.246
2. % managerial workers	.609	.371	.125
3. % Russian population	.640	.410	.039

The biggest surprise was the strong correlation between the occupational and local ambient environment indicators and female breast cancer mortality. When the VMT/wind direction/ distance indicator was inserted as a surrogate for over 100 significant occupation and local ambient environment indicators, it accounted for one-third of the variance. The indicator of socioeconomic status was reduced in explanatory power from more than 20 percent when the ethnic, occupational and environmental variables were not included (Table 64) to thirteen percent when they are included (Table 66). The VMT index accounts for between 32 and 59 percent of the variance.

TABLE 66

Multivariate Analysis of White Female Breast Cancer Mortality, 1950-1969, All Significant Indicators

Variable	45 Areas Mult. R	45 Areas R^2	45 Areas Change R^2	21 Areas Mult. R	21 Areas R^2	21 Areas Change R^2
1. VMT/wind direction/ distance index	.573	.328	.328	.767	.588	.589
2. % managerial workers	.677	.458	.130	-	-	-
3. % Italian population	.706	.498	.040	-	-	-
4. % German population	.732	.536	.038	-	-	-
5. Female median education	-	-	-	.821	.674	.086
6. % Polish population	-	-	-	.839	.704	.030

The final multivariate analysis used the three statistically created and uncorrelated variables. These conservative indicators place the contribution of socioeconomic status at 17 percent and ethnic, occupational, and local ambient environmental factors at 31 percent (Table 67).

Table 67

Multivariate Analysis of White Female Breast Cancer Mortality, 1950-1969, Factor Analysis Variables

45 Areas

Variable	Mult R	R^2	Change R^2
1. Urban/industrial Core	.512	.262	.262
2. Socioeconomic Status	.657	.432	.170
3. Polish/Italian Corridor	.692	.479	.047

Summarizing, the expected positive correlations between white female breast cancer mortality and socioeconomic status materialized. They accounted for between 8 and 35 percent of the variance. Unexpected ethnic correlations accounted for between 3 and 29 percent, when introduced. Finally, when introduced, occupational and local ambient environment indicators accounted for between 30 and 59 percent of the variation in white female breast cancer mortality. The results clearly call for further analyses. Suggested epidemiological studies include the Italian, USSR, Polish, German, and native-born populations in counties such as Essex, Union, Sussex and Cumberland Counties. If at all possible, foreign born and first generations should be compared to later generations of the same foreign stock. In addition, records of cervix uteri cancer mortality should be gathered for the same populations to provide a control disease that should be expected associate negatively with female breast cancer.

White Female and Male Lung Cancer Mortality

Lung cancer is the major cause of cancer related death among men. More than 20 percent of white male cancer mortality deaths result from neoplasms of the lung and trachea. Among the female population lung cancer is responsible for less than 5 percent of cancer mortalities. Male cancer mortality rates in the region are more than six times female rates.

While male rates are much higher, the two tend to covary geographically. One of five counties with the highest rates and three of five with the lowest rates of lung cancer mortality are common to both the male and female population. The correlation between white female and male lung cancer mortality was .571. The male lung cancer mortality rate in the region was 12 percent above the national average, the female rate was 10 percent above the national average. Both exceed the 80th percentile for all the states. Overall, while male lung cancer mortality rates are six times female rates, the spatial distribution of the male and female rates is relatively similar.

We expected that the high risk factors would be similar: smoking, ethnic including higher rates for Polish, USSR, and to a lesser extent German populations, low socioeconomic status, and air pollution. The additional important high risk factor for the male population is occupation. In the absence of cigarette smoking data, we had to assume that smoking, if a factor, was cocarcinogenic along with the other high risk factors.

Thirty of 174 correlations between white female lung cancer and high risk factors were significant (Table 68).

TABLE 68

Important Bivariate Correlations: White Female Lung Cancer Mortality, 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R
No. of Significant Correlations	24	4	2
1. % Italian population	.400*	.366*	-
2. TSP/wind direction/distance index	.359*	.291	-
3. VMT/wind direction/distance index	.318	-	-
4. Manufacturing workers/distance index	.299	-	-
5. Control variable: population density	.299	-	-
6. % Female married	-	-	-.375
7. % USSR population	-	-	.373
8. Urban/industrial core factor	.383*	a	a

* Significant at .01 level
a Only considered for n=45 case

No indicator is consistent across the three geographical regions. The Italian variable and air pollution index appear in the 45 and 40 case regions. The Italian correlation was unexpected, the others were expected. The stepwise multivariate analysis accepted the Italian and TSP index for the 45 case region and the female married indicator for the 21 case region (Table 69). Perhaps, the female married correlation is an indication of female smoking habits. Only 21 and 14 percent of the variance are explained, respectively, in the 45 and 21 case regions.

TABLE 69

Multivariate Analysis of White Female Lung Cancer Mortality, 1950-1969

Variable	45 Areas Mult. R	45 Areas R^2	45 Areas Change R^2	21 Areas Mult. R	21 Areas R^2	21 Areas Change R^2
1. % Italian population	.400	.160	.160	-	-	-
2. TSP/wind direction/ distance index	.462	.213	.053	-	-	-
3. % female married	-	-	-	.375	.141	.141

Frankly, the statistical analyses of white female lung cancer mortality were weaker than anticipated. An examination of the highest rates suggests unexplained factors. The high rates in New York City and Nassau County may be explained by the indicators we have used. The high rates for Putnam, Somerset, Monmouth, and Monroe Counties require further thought and research. Putman County's rate was 10.5/100,000 a rate almost double that of Dutchess County which lies immediately north of Putnam. Could Putnam's rate be accounted for by migration from New York City? Monroe County's high rate is inexplicable to us. Perhaps smoking and population migration are among the crucial factors that might account for these anomalies.

Overall, the high risk indicators help us understand the geographical distribution of low risk counties, but do not satisfactorily explain the high risk counties. One clear epidemological study is suggested: a comparison of white female lung cancer mortality cases in Putnam and Dutchess Counties, New York.

The analysis of white male lung cancer mortality was more complex and revealing than the similar study of the female rates. The five highest rates are in or adjacent to the region's urban/industrial cores (Hudson, Middlesex, Camden, and Nassau Counties and Philadelphia); the lowest rates are found in relatively rural counties in the northern and western portion of the study area (Carbon, Chester, Lehigh, Litchfield, Dutchess, and Monroe, see Chapter 2 for a map). Differences in lung cancer mortality rates between the high and low counties are substantial. Hudson County's 1950-69 rate is 57.4/100,000; Carbon's is 29.9.

The 126 significant bivariate correlations will be reviewed in detail because they contain an unclear and subtle message about the role of occupational exposure in inducing white male lung cancer mortality. The message is that the role of industry, especially the chemical industry, and the ethnic groups are relatively more important than air pollution when the untransformed indicators are used. When the transformed indicators are used the relative importance of air pollution increases (Table 70).

TABLE 70

Important Bivariate Correlations: White Male Lung Cancer, 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R
No. of Significant Correlations	46	38	42
1. Control variable: population density	.427*	.528*	.575*
2. Manufacturing workers	.250	.316	.522*
3. Chemical workers	.459*	.419*	.472
4. Construction workers	.254	.344	.412
5. TSP	.330	.420*	.603*
6. TCE	.456*	.519*	.695*
7. % employed in transportation equipment manufacturing	.308	.284	-
8. % Polish population	.456*	.473*	.590*
9. % Italian population	.483*	.439*	.489
Land Area Indices			
10. Manufacturing workers/land area index	.467*	.485*	.581*
11. Chemical workers/land area index	.508*	.459*	.531*
12. Construction workers/land area index	.409*	.505*	.506*
13. TSP/land area index	.504*	.567*	.660*
14. TCE/land area index	.440*	.486*	.639*
Wind direction/distance and distance indices			
15. Manufacturing workers/distance index	.566*	.541*	.576*
16. Chemical workers/distance index	.552*	.513*	.528*
17. Construction workers/distance index	.592*	.567*	.588*
18. TSP/wind direction/distance index	.520*	.645*	.691*
19. TCE/wind direction/distance index	.569*	.567*	.706*

TABLE 70 (continued)

Important Bivariate Correlations: White Male Lung Cancer, 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R
No. of Significant Correlations	46	38	42
20. VMT/wind direction/distance index	.392*	.389*	.602*
21. Urban/Industrial core factor	.523*	a	a
22. Polish/Italian corridor factor	.328	a	a

* Significant at .01 level
a Considered only for n=45 area

Variables 1-9 (Table 70) are not transformed by wind direction and distance. Were we to examine the entire study area (45 cases) we would conclude that the relative presence of an Italian population (R=.483) and the presence of chemical workers (R=.459) are the most important high risk factors, with the Polish population and the burning of fossil fuels in electricity generating stations close behind. These four indicators have higher correlations than the control variable population density.

When the employment and air pollution indicators are divided by land area (variables 10-14), chemicals and particulates become the most significant indicators. Finally, when the distance and wind direction/distance indices are calculated (variables 15-20), construction workers and fossil fuels become the most important indicators.

Switching of the relative importance of different high risk factors was unusual for this study. Normally, the strongest correlations produced by untransformed data remained the strongest coarrelations after transformation. We used the transformed indices in the subsequent correlation analysis because they yielded the strongest associations. For consistency, we followed the same procedure in the multivariate analyses of white male lung cancer mortality.

Before reviewing the multivariate analyses, two observations will be made. First, we had expected strong correlations between white male lung cancer and the presence of the construction, chemical, and metal industries. Using all manufacturing workers and population density as control variables, chemical workers and to a lesser extent construction and motor vehicle workers demonstrated strong correlations. The second observation is that while we believe that these occupational and ethnic correlations are real, they are overshadowed by the transformed air pollution indices in the multivariate analyses. They do not appear because the transformed air pollution indices produced higher correlations than the transformed occupational variables and are therefore chosen by the stepwise program.

The occupational and ethnic variables also do not appear because they are highly intercorrelated with each other as demonstrated below (Table 71).

TABLE 71

Correlations of Selected High Risk Indicators, 45 Area Region

	Variable							
Variable	1	2	3	4	5	6	7	8
1. TCE/wind direction/distance index	-							
2. % Italian population	.441*[a]							
3. % Polish population	.525*	.531*						
4. Chemical workers/distance index	.775*	.440*	.639*					
5. Manufacturing workers/distance index	.875*	.509*	.592*	.964*				
6. Construction workers/distance index	.867*	.513*	.578*	.952*	.996*			
7. % transportation equipment workers	-.016	.270	.173	.042	.033	.049		
8. Population density	.783*	.361*	.369*	.688*	.783*	.800*	-.027	

* Significant at .01 level

[a] .441 is the correlation between variables land 2

The degree of intercorrelation among the high risk lung cancer indicators makes it infeasible to pull apart the separate contributions of the personal, occupational, and local ambient risk factors. Twenty-one of the 28 bivariate correlations in Table 71 are significant at the .01 level. The only insignificant correlations are between the percentage of transportation equipment workers and the other seven variables. Several path analyses were designed and rejected because of a lack of confidence in our ability to build the proper paths.

Given the high degree of intercorrelation among almost all of the important independent variables, the simple, stepwise multivariate analysis results must be interpreted with the highest degree of caution. The results literally mean that the combustion of fossil fuels for electricity is strongly correlated with white male lung cancer (Table 72). The presence of motor vehicle and chemical manufacturing and of an Italian population are of lesser significance. Realistically, we know that the TCE/wind direction/ distance index is a surrogate and strongly correlated with the TWP/wind direction/distance index (R=.943). the occupational, and even the ethnic factors. In short, we cannot clearly distinguish among the high risk components for white male lung cancer because of the limitations of the mathematical methods at the geographical scale of the analysis and perhaps because they may operate synergistically in the real world.

TABLE 72

Multivariate Analysis of White Male Lung Cancer
Mortality, 1950-1969

Variable	45 Areas			45 Areas		
	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2
1. TCE/wind direction/ distance index	.569	.324	.324	.706	.498	.498
2. % motor vehicle workers	.651	.424	.100	.765	.586	.088
3. % Italian population	.673	.453	.029	-	-	-
4. Chemical workers/ distance index	.684	.468	.015	-	-	-
5. % Polish population	-	-	-	.785	.616	.030
6. Urban/industrial core factor	.523	.273	.273	a	a	a
7. Polish/Italian corridor factor	.612	.375	.102	a	a	a

a Computed only for n=45 region

The simple stepwise procedure accounts for almost 47 percent of variance in the 45 county region and more than 60 percent in the New Jersey study area. The urban/industrial core and Polish/Italian corridor factors are not correlated. These two factors accounted for about 38 percent of the variance in white male lung cancer mortality. The explanatory power of the two factors implies, in turn, that the core and corridor counties have the highest white male lung cancer rates.

The essence of the results of the analysis of white male lung cancer mortality has been to more clearly define for our study area the constitutents of the "urbanization" and "English" factors cited in the literature. While we do not have a clear answer, we may conclude that with the exceptions of most of the metals industries, the USSR population, and income, the expected associations materialized. In addition, the analysis leads to suggested research. Why were correlations with the Italian population unexpectedly high, while associations with the USSR population variable lower? Are occupation, place of residence, air pollution, smoking, or other factors implicated? Three types of studies are suggested. One would involve a comparison of two or more ethnic groups (USSR, Italian, Polish) in a county that has a large representation of these ethnic groups (e.g., Essex and Union counties.)

The second is a comparison of nationality groups between counties. For example, Hudson and New Haven Counties have about the same proportion of Italian, USSR, and Polish populations. Hudson County's lung cancer rate is 57.4/100,000 and New Haven's 41.6. Nassau and Litchfield Counties have similar proportions of Italian populations. Nassau's lung cancer rate is 50/100,000, Litchfield's 35.2/100,000.

A third suggested comparison is between the industrial structure of counties with high and low lung cancer rates. In comparison to Salem County, four counties (Somerset, New Castle, Middlesex and Passaic) had relatively low bladder cancer, but relatively high white male lung cancer rates. Can these differences be traced to occupational factors (for example, asbestos production in Somerset and Passaic)? Epidemiological workforce studies and product inventory analyses would be required. Finally, we are trying to repeat the analysis with a surrogate for smoking habits as a control for the Part II study. The parts II and III studies will also contain interesting comparisons of the change in lung cancer during the 1950's and 1960's.

White Female and Male Large Intestine Cancer Mortality

Cancer of the large intestine is a major disease in the United States, and especially in the study area. During the 1950-1969 period, it was the second leading cause of cancer mortality among both females and males. More than 10 percent of white male and about 14 percent of white female cancer mortalities were attributed to the large intestine. The female and male rates are respectively 24 and 23 percent higher than the national averages. The region's rates exceed the upper decile for all the states. In short, cancer of the large intestine is a disease of great importance in the study area.

A set of high risk factors matching the importance of the disease was not available. Diet is strongly suspected. Low association with USSR, German, and Italian (male only) populations were expected. No association with the Polish variable was expected. We further assumed, albeit without much support from the literature, that local ambient, environment and occupational exposures might be linked. Overall, we expected that male and female high risk factors would be similar.

The results did not support the link between male and female high risk factors and, in general, were extremely difficult to interpret. A review of the high and low rate counties immediately suggested that the expected associations would not materialize. The five high rate male counties are along the urban/

industrial core: Philadelphia, Essex, New York City, Hudson, and Camden. Camden and Philadelphia are common to the high male and female group. However, the other high rate counties among the female population are Pike, Hunterdon, and Burlington. While the high male rates are centered in the region's urbanized cores, the high female rates are in urban and rural places in the Delaware River Valley. Indeed, Pike and Hunterdon Counties were among the high female counties and low male counties. Overall, it was not surprising that the correlation between male and female large intestine cancer mortality was insignificant.

White female large intestine cancer mortality produced 41 significant bivariate correlations (Table 73).

TABLE 73

Important Bivariate Correlations: White Female Large Intestine Cancer Mortality, 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R
No. of Significant Correlations	23	15	3
1. Control variable: population density	.264	.268	-
2. manufacturing workers/distance index	.264	.268	-
3. TSP/wind direction/ distance index	.281	.301	-
4. % USSR population	.252	.251	-
5. % self-service water supply	-	-.260	-.423
6. % female married	-	-	-.309
7. Urban/industrial core factor	.317	a	a

a Only considered for n=45 case region

The 41 bivariate correlations do not include a single correlation significant at the .01 level. The 41 correlations imply that in general, the urban/industrial core counties have slightly higher rates than the other counties. We can conclude little else from the 45 and 40 case studies. The analysis of New Jersey's 21 counties implies that water supply may be implicated that married women have lower rates. Both of these correlations are tenuous because they only hold for New Jersey. The white female correlations are so weak that multivariate analysis was considered inappropriate. In conclusion, the white female large intestine results were surprisingly weak.

The analysis of white male cancer mortality produced 133 significant correlations. Many were significant at the .01 level. The most important bivariate and multivariate correlations are summarized below (Tables 74 and 75).

TABLE 74

Important Bivariate Correlations, White Male Large Intestine Cancer Mortality, 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R
No. of Significant Correlations	49	44	40
1. Control variable: population density	.484*	.465*	.481
2. manufacturing workers/distance index	.392*	.481*	.573*
3. TSP/wind direction/ distance index	.539*	.512*	.550*
4. % Italian population	.514*	.451*	.490
5. % self-service water supply	-.412	-.297	-.416
6. % Polish population	.391*	.319	-
7. Urban/industrial core factor	.502*	a	a
8. Polish/Italian corridor factor	.271	a	a

* Significant at .01 level

a Only considered for n=45 case

TABLE 75

Multivariate Analysis of White Male Large Intestine Cancer Mortality, 1950-1969

	45 Areas			21 Areas		
Variable	Mult. R	R^2	Change R^2	Mult. R	R^2	Change R^2
1. TSP wind direction/ distance index	.539	.291	.291	-	-	-
2. % Italian	.639	.408	.117	-	-	-
3. Manufacturing workers	-	-	-	.573	.329	.329
4. Urban/industrial	.502	.252	.252	a	a	a
5. Polish/Italian corridor factor	.566	.320	.068	a	a	a

a Only considered for n=45 case region

The correlations are strong, but frankly questionable. The Polish population variable was not expected to correlate; it did. The correlation of the air pollution and manufacturing indicators to the best of our knowledge has no basis in the literature, though one might speculate on possible links. Perhaps the correlations signify an increased risk; however, our conservative guidelines prevent the immediate acceptance of these correlations.

The Italian correlation was expected and the water supply indicator may be real. Our confidence in these correlations is not great. In conclusion, we observe that white male large intestine cancer mortalities tend to be higher in the urban/industrial core and ethnic corridor counties. However we could not sort out the high risk factors. The analysis does suggest that the male Italian ethnic population in New York City, Hudson, Essex, Philadelphia, and Camden Counties is at high risk. The water supply link is interesting because it was consistent across the male and female populations. Could there be a link between cancer of the large intestine and water supply? Counties that have a large proportion of their population served by public potable water supply systems have higher large intestine cancer rates. Whether the correlation is real remains to be seen and is a good research topic.

White Female and Male Liver Cancer Mortality

Liver cancer is responsible for about three percent of the region's cancer mortalities. The regional rates are about two percent higher than the national average. Female mortality rates are slightly higher than male rates (5.19 for white males, 5.52 for white females).

Smoking and alcohol consumption have been implicated in liver cancer. While we do not have smoking and alcohol data at this time, we note no significant correlation between white female lung cancer and white female liver cancer. However, a significant correlation between white male liver and lung cancer mortality was found (R=.500). Accordingly, if lung cancer is a surrogate for smoking (a fact about which we have some doubts in the study area), then we cannot dismiss smoking as a factor.

While our analysis lacks all the necessary risk factors, important risk factors are present in the region. In particular among the female population, we expected to find negative associations with socioeconomic status. Among the male population, the socioeconomic status factor was expected to be joined by industrial and ethnic correlations. The presence of the chemical industry was expected to produce a strong positive association. Weaker positive associations were expected with the textiles and metals industries. Finally, weak positive associations were expected between the white male rate and the presence of Italian populations and between the Polish population and both the male and female rates. Positive, albeit weak, associations were also thought possible with air and water quality.

The results consistently and strongly supported most of the hypothesized associations. Forty-nine of the 174 possible female bivariate correlations were significant. Seven indicators were consistently strong (Table 76).

TABLE 76

Important Bivariate Correlations: White Female Liver Cancer Mortality, 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R
No. of Significant Correlations	12	28	9
1. % female completed at least college	-.478*	-.533*	-.534*
2. % professional workers	-.464*	-.515*	-.472
3. Female median education	-.452*	-.474*	-.488
4. % managerial workers	-.424*	-.290	-.618*
5. Median income	-.297	-.267	-
6. TCE/wind direction/distance index	-.277	.359	.468
7. Socioeconomic status factor	-.459*	a	a

* Significant at .01 level
a Only considered for n=45 case

Six of the seven are the expected socioeconomic status indicators. The seventh is the fossil fuel emissions indicator which we interpret as an air emissions indictor. The multivariate analysis accepted the most significant socioeconomic status indicator and the TCE/wind direction/distance index (Table 77).

TABLE 77

Multivariate Analysis of White Female Liver Cancer Mortality, 1950-1969

Variable	45 Areas Mult. R	R^2	Change R^2	21 Areas Mult. R	R^2	Change R^2
1. % female completed at least college	.478	.229	.229	-	-	-
2. % managerial workers	-	-	-	.618	.382	.382
3. TCE/wind direction/distance	.524	.274	.045	.692	.478	.096

It seems clear that relative affluence is strongly associated with white female liver cancer mortality. The air quality indicator is less strong. Is affluence a surrogate for diet, smoking and alcohol consumption patterns? This study cannot offer an answer. However it has suggestions about where to begin the search. The highest rates were in Carbon, Monroe, Hudson, and Middlesex Counties; the lowest in Pike, Rockland, Dutchess, Orange, and Suffolk Counties. Carbon County, Pennsylvania has the highest white female cancer mortality rate in the study area. When this observation is combined with the fact that Carbon had the lowest breast cancer mortality, we conclude that Carbon is a prime place to study female cancer mortality patterns.

The correlation between white female and male liver cancer mortality is only .280. Therefore, we did not expect socioeconomic status to dominate the male analysis. Three fourths (129) of the bivariate correlations between white male liver cancer and the high risk factors were significant. The correlations of the important 18 variables are indicated below (Table 78).

TABLE 78

Important Bivariate Correlations: White Male Liver Cancer Mortality, 1950-1969

Variable	45 Areas R	40 Areas R	21 Areas R
No. of Significant Correlations	47	37	45
1. Control variables: manufacturing workers/distance index	.610*	.571*	.607*
2. Manufacturing workers/land area index	.534*	.525*	.617*
3. Manufacturing workers	.363*	.404*	.660*
4. Population density	.492*	.540*	.640*
Untransformed Indicators			
5. % Polish population	.436*	.374*	.471
6. % Italian population	.440*	.402*	.504*
7. % self-service water supply	-.384*	-.258	-.552*
8. TCE	.486*	.450*	.640*
9. TSP	.321	-	.561*
10. Chemical workers	.580*	.550*	.623*
Area Index Indicators			
11. TCE/land area	.464*	.435*	.531*
12. TSP/land area	.495*	.467*	.665*
13. Chemical workers/land area	.581*	.552*	.634*
Distance and Wind Direction/Distance Variables			
14. TCE/wind direction/distance index	.569*	.525*	.615*
15. TSP/wind direction/distance index	.507*	.522*	.692*
16. Chemical workers/distance index	.613*	.582*	.593*
Factor Analysis Variables			
17. Urban/industrial core factor	.488*	a	a
18. Polish/Italian corridor factor	.364*	a	a

* Significant at .01 level
a Only considered for n=45 area

The income hypothesis is rejected. However, some of the industrial, ethnic, and local ambient risk factors appear. In particular, the chemical workers correlation is consistent and high. The chemical correlations are considerably higher than the other manufacturing industries and population density in the 45 case region, though not for the New Jersey study area. As in the case of the white male lung cancer mortality correlations, we cannot point to a unique male liver cancer-chemical industry relationship. The high risk factors are too highly intercorrelated for such a simple conclusion. Frankly, however, the strong contrast between the factors driving the male and female correlations strongly suggests occupational and to a lesser extent local ambient environmental factors. In the lung cancer case, a similar conclusion could not be drawn because the correlation between male and female white lung cancer was .571, suggesting some similar causal factors, while as noted above the correlation between white male and female liver cancer is only .280.

The stepwise multivariate results must be interpreted in light of the above conclusions. The chemical/distance index indicator is selected in the 45 case region and the TSP air pollution index in the 21 case New Jersey region (Table 79). The two statistically derived factors account for more than one-third of the variance of white male liver cancer mortality. The more specific chemical and TSP driven multivariate analyses covary to a larger degree.

TABLE 79

Multivariate Analysis of White Male Liver Cancer Mortality, 1950-1969

Variable	45 Areas Mult. R	R^2	Change R^2	21 Areas Mult. R	R^2	Change R^2
1. Chemical workers/ distance index	.613	.376	..376	-	-	-
2. % Italian population	.641	.411	.035	-	-	-
3. TSP/wind direction/ distance index	-	-	-	.692	.479	.479
4. % Russian population	-	-		.747	.558	.079
5. Urban/industrial core factor	.488	.238	.238	a	a	a
6. Polish/Italian corridor factor	.604	.365	.127	a	a	a

a Only considered for n=45 case region

The most obvious research project suggested by the liver cancer analysis is a comparison of industrial chemicals and industrial workers in two sets of counties. Hudson, Middlesex, and Essex Counties have higher liver mortality rates (6.7-7.1), while Salem, Somerset, and Union Counties have lower white male liver cancer rates (4.3, 4.5 and 5.8, respectively). Overall, despite the relative lack of importance of liver cancer as a cause of mortality, it appears to be a prime disease for investigation.

White Female and Male Rectum Cancer Mortality

Mortality caused by cancer of the rectum is much higher in the study area than in the nation as a whole. The region's white male rate is 36 percent higher than the national average. It exceeds the upper decile for the states. The study area's white female rate is 28 percent higher than the national average and is higher than the rate for 80 percent of the states. Overall, cancer of the rectum is responsible for about five percent of white male cancer mortalities and four percent of white female cancer related mortalities in the region.

The literature offers few high risk factors. The only literature grounded hypotheses were ethnic. Based on dietary habits, we expected to find positive associations with the Polish, USSR, and Italian variables. We assumed, although perhaps without justification, that cancerous conditions could be initiated and/or sustained in the rectum due to occupational and local ambient environmental conditions. Overall, we expected to find relatively few significant correlations.

In contrast to what we had expected, we found many significant correlations. Initial hints of what was to happen were given by the .638 correlation between male and female rectum cancer mortality, and equally significant correlations between rectum, stomach, and all cancers. More than 250 significant bivariate correlations were recorded. None stand out; most are of moderate strength (between R=.40 and .50).

The population density and manufacturing control variables produced higher correlations than the expected Italian, USSR, and Polish variables (Table 80).

TABLE 80

Important Bivariate Correlations: White Female and Male Rectum Cancer Mortality, 1950-1969

Variable	45 Areas R		40 Areas R		21 Areas R	
	WF	WM	WF	WM	WF	WM
No. of Significant Correlations	45	46	35	42	42	43
1. Control variable: population density	.435*	.483*	.500*	.571*	.584*	.632*
2. Manufacturing workers/ distance index	.518*	.579*	.523*	.561*	.546*	.647*
3. % Italian population	.443*	.368*	.501*	.411*	.454	.473
4. TCE/wind direction/ distance index	.436*	.482*	.410*	.436*	.472	.496*
5. % females married	-.417*	NR	-.292	NR	-.525*	NR
6. % self-service water	-.397*	-.344*	-.383*	-.405*	-.487	-.337
7. Median education	.302	-	.291	-	.443	-
8. % Polish population	.264	.416*	-	.441*	-	.478

* Significant at .01 level
NR= Not relevant

The combinations of highly interdependent independent variables and weak high risk hypotheses led to unsatisfactory multivariate, stepwise regression results. The 45, 40, and 21 case results were extremely different.

Accordingly, we fell back to the uncorrelated variables produced by the factor analysis (Table 81).

TABLE 81

Multivariate Analysis of White Female and Male Rectum Cancer Mortality, 1950-1969

Variable	45 Areas					
	White Female			White Male		
	Mult R	R^2	Change R^2	Mult R	R^2	Change R^2
1. Polish/Italian corridor factor	.396	.157	.157	.443	.197	.197
2. Urban/industrial core factor	.529	.280	.123	.595	.354	.157

The two factors imply a moderately strong association between the high risk ethnic groups, urbanization/industrialization, public water supplies, air pollution and female and male rectum cancer mortality. The common high mortality counties among the female and male population are Hudson, Essex, Passaic, and New York City. The low mortality counties are inconsistent.

Overall, we have found moderately strong correlations between many of the high risk indicators and rectum cancer. However, the relative importance of specific indicators could not be determined. We suspect that occupational factors are not important because the male and female rectum rates are strongly associated. The positive ethnic correlations suggest personal environmental, particularly dietary factors. However, unlike the case of white male leukemia where we could argue for personal environmental factors because the ethnic correlations were clearly the highest, in this case, the personal, occupational and local ambient factors are about equally associated with rectum cancer. Therefore, while we suspect dietary factors, this analysis did not clearly distinguish them.

White Female and Male Stomach Cancer Mortality

During recent decades, stomach cancer rates have decreased. Nevertheless, the disease is the third leading cause of cancer mortality among men in the region, accounting for nine percent of white male cancer mortalities. Women's rates are about half of men's. The region's rates are significantly higher than the nation's. The regional male rate exceeds the national rate by 12 percent and is above the eightieth percentile for the states. The female rate exceeds the national average by 13 percent.

We expected and found strong associations. High positive associations were expected with Polish and USSR populations. Lower correlations with German and Italian nationals also were expected. Socioeconomic status was expected to correlate negatively. Smoking could not be ruled out as a factor because male and female rates were significantly correlated with lung cancer. Finally, occupational and local ambient environmental factors were expected to correlate.

Stomach cancer mortality produced more significant correlations than any other body site. A total of 280 significant bivariate correlations were recorded, half with male and half with female. These results suggest that, as in the case of rectum cancer, we would not be able to separate the interacting high risk factors.

In this case, however, one indicator is much stronger than any other: the Polish population.

The male and female rates are tabulated adjacent to one another in Table 82. The correlation between female and male stomach cancer was .835, the highest correlation between two cancers. This correlation suggests common causal factors.

TABLE 82

Important Bivariate Correlations: White Female and Male Stomach Cancer Mortality, 1950-1969

Variable	45 Areas R		40 Areas R		21 Areas R	
	WF	WM	WF	WM	WF	WM
No. of Significant Correlations	50	51	44	43	46	46
1. % Polish population	.717*	.733*	.771*	.742*	.851*	.816*
2. manufacturing workers/ distance index	.606*	.585*	.610*	.542*	.730*	.638*
3. TCE/wind direction/ distance index	.584*	.597*	.563*	.547*	.641*	.636*
4. control variable: population density	.509*	.495*	.566*	482*	.664*	.601*
5. TSP/wind direction/ distance index	.488*	.469*	.527*	.409*	.785*	.744*
6. % surface water supply	.425*	.398*	.390*	.323	.796*	.703*
7. % USSR population	.424*	.440*	.411*	.363	.507*	.449
8. % Italian population	.414*	.434*	.504*	.456*	.657*	.688*
9. Urban/industrial core factor	.532*	.507*	a	a	a	a
10. Polish/Italian corridor factor	.479*	.502*	a	a	a	a
11. socioeconomic status factor	–	-.346	a	a	a	a

* Significant at .01 level

a Only considered for n=45 area

The stepwise multivariate analysis chose the Polish variable first in all runs. Industrial activity and Italian population were selected as secondary factors (Table 83).

TABLE 83

Multivariate Analysis of White Female and Male Stomach Cancer Mortality, 1950-1969

Variable	45 Areas			21 Areas		
	Mult R	R^2	Change R^2	Mult R	R^2	Change R^2
A. White Male						
1. % Polish population	.733	.537	.537	.816	.666	.666
2. All manufacturing employees/land area index	.780	.609	.072	-	-	-
3. % Italian population	-	-	-	.859	.737	.071
4. Urban/industrial core factor	.507	.257	.257	a	a	a
5. Polish/Italian corridor factor	.707	.500	.243	a	a	a
6. Socioeconomic status factor	.786	.618	.118	a	a	a
B. White Female						
1. % Polish population	.717	.514	.514	.851	.724	.724
2. Manufacturing workers/distance index	.768	.589	.075	-	-	-
3. Manufacturing workers	-	-	-	.907	.823	.099
4. Urban/industrial core factor	.532	.283	.283	a	a	a
5. Polish/Italian corridor factor	.709	.503	.220	a	a	a

a Only considered for n=45 area region

If the Polish variable correlation as a surrogate for dietary factors is a direct link to stomach cancer, then state government policy options are limited. Indeed, if stomach cancer is driven by imported dietary habits, as our analysis suggests, then stomach cancer should continue to decline as dietary habits shift away from those of the country of orgin.

On the basis of the analysis, further studies of stomach cancer should be focused on the eating habits of Polish and to a lesser extent Italian and Russian Americans. The counties of major interest follow the urbanized spine of New Jersey: Hudson, Passaic, Union, Middlesex, and Mercer. Hudson and Middlesex are particularly recommended as study areas. The life styles of Chester and Montgomery County citizens should offer (with their very low rates) a strong constrast with respect to high risk stomach cancer mortality factors.

White Female and Male All Cancer Mortality

The study area's female and male cancer mortality rates are ten percent above the national average and exceed the upper decile for all the states. The specific cancer diseases that have been reviewed account for 66 percent of the female and 77 percent of the male mortalities. Accordingly, we should expect the correlations found with lung, stomach, large intestine, and some of the other major body sites to be replicated in the analysis of cancer deaths at all sites combined.

The best way of summarizing the results is to tabulate the highest correlations in each high risk category (Table 84).

TABLE 84

Important Bivariate Correlations, White Female and Male All Cancer Mortality, 1950-1969

Variable	45 Areas R WF	45 Areas R WM	21 Areas R WF	21 Areas R WM
1. Control variable: population density	.497*	.606*	.666*	.762*
Personal Environment				
2. % Polish population	.257	.583*	-	.641*
3. % USSR population	.312	.430*	-	.412*
4. % Italian population	.254	.573*	.472	.679*
5. % same house	.270	-	.481	-
Occupational Environment				
6. Manufacturing employees/ distance index	.478*	.691*	.588*	.725*
7. % agricultural workers	-.317	-	-.740*	-.422
Local Ambient Environment				
8. TSP/wind direction/distance index	.494*	.654*	.669*	.842*
9. % self-service water supply	-	-.500*	-.500*	-.639*
Factor Analysis Variables				
10. Urban/industrial core factor	.530*	.633*	a	a
11. Polish/Italian corridor factor	-	.454*	a	a

* Significant at .01 level

a Only considered for n=45 case

b 40 cases not presented because they are redundant

Table 84 once again suggests that population density is a good surrogate for the spatial distribution of all cancer. The manufacturing occupation category variable yields the highest correlation for the white male population 45 case region, the urban/industrial core factor yields the highest correlation for the female population.

The multivariate analysis (Table 85) of all cancers must be interpreted with great caution. The high correlation with the manufacturing/distance index undoubtedly signifies not only occupational, but also local ambient and even personal environmental factors. Population density and air pollution play the same role in other white male and white female multiple regressions. The urban/industrial core and Polish/Italian corridor factors strongly correlate with white male all cancer mortality. The urban/industrial core factor strongly correlates with white female cancer mortality. In short, the all cancer relationships underscore the high rates and some of the reasons why high cancer rates are found in Hudson, Philadelphia, Middlesex, and Essex Counties and New York City.

TABLE 85

Multivariate Analysis of White Female and Male All Cancer Mortality, 1950-1969

A. White Female

Variable	45 Areas Mult R	45 Areas R^2	45 Areas Change R^2	21 Areas Mult R	21 Areas R^2	21 Areas Change R^2
1. Population density	.497	.247	.247	-	-	-
2. % professional workers	.552	.305	.058	-	-	-
3. Manufacturing workers/ distance index	.587	.345	.039	-	-	-
4. TSP/wind direction/ distance index	-	-	-	.669	.448	.448
5. % different house, same county	-	-	-	.803	.645	.197
6. Urban/industrial core factor	.530	.280	.280	a	a	a
B. White Male						
1. Manufacturing workers/ distance index	.691	.478	.478	-	-	-
2. % self-service water	.740	.548	.070	-	-	-
3. % Russian population	.764	.584	.036	-	-	-
4. % Italian population	.785	.616	.032	-	-	-
5. TSP/wind direction/ distance index	-	-	-	.842	.709	.709
6. % agricultural workers	-	-	-	.886	.785	.076
7. Urban/industrial core factor	.633	.400	.400	a	a	a
8. Polish/Italian corridor factor	.772	.596	.196	a	a	a

a Only considered for n=45 area

Urban/Industrial Cancer Corridor Factor

The second chapter described 14 counties that manifested high and 11 counties that had low cancer rates of all sites, stomach, male large intestine, rectum, lung, female breast, and male liver cancer mortality. The unrotated factor loading and factor scores of the strongest cases are tabulated below (Tables 86 and 87). (Both the unrotated and rotated scores are found in Chapter 2).

TABLE 86

Urban-Industrial Cancer Corridor Factor

Variable	Unrotated Factor Loadings
1. WM all sites	.946
2. WF all sites	.894
3. WF stomach	.789
4. WM stomach	.764
5. WM large intestine	.722
6. WM rectum	.700
7. WM lung	.689
8. WM liver	.683
9. WF breast	.659
10. WF rectum	.653
11. WF lung	.493

TABLE 87

Counties Belonging to Urban/Industrial Cancer Corridor Group

County	High Cancer Unrotated Factor Score	County	Low Cancer Unrotated Factor Score
1. Hudson	2.20	Dutchess	-2.14
2. New York City	2.03	Chester	-2.03
3. Philadelphia	1.67	Suffolk	-1.47
4. Middlesex	1.56	Bucks	-1.36
5. Essex	1.50	Montgomery	-1.14
6. Nassau	1.40	Rockland	-1.06
7. Passaic	1.15		
8. Union	1.10		

The cancer corridor factor is strongly correlated with the all white female and male cancer categories. Accordingly, the highly interdependent personal, occupational and local ambient variables are represented in multivariate analysis by a few simple surrogates (Table 88).

TABLE 88

Multivariate Analysis of Urban/Industrial Cancer Corridor Factor

Variable	45 Areas Mult R	 R^2	 Change R^2
1. Manufacturing workers/distance index	.640	.410	.410
2. TSP/wind direction/distance index	.749	.561	.151
3. Urban/industrial core factor	.654	.428	.428
4. Polish/Italian corridor factor	.737	.544	.116

The Southern New Jersey Factor

Nine counties were marked by high rates of female and male bladder cancer, female large intestine and male leukemia cancer. The varimax rotated factor loadings and average of the unrotated and varimax rotated factor loadings are summarized below (Tables 89 and 90). (The unrotated and rotated solutions are found in Chapter 2).

TABLE 89

White Population Southern New Jersey Cancer Factor

Variable	Varimax Rotated Factor Loading
1. Female bladder	.831
2. Female large intestine	.688
3. Male bladder	.626
4. Male leukemia	.484

TABLE 90

Factor Scores for Counties Strongly Identifying With Southern New Jersey Factor

County	Average of Rotated and Unrotated Factor Score
1. Pike	3.92
2. Salem	1.89
3. Gloucester	1.29
4. Philadelphia	0.85
5. Nassau	0.85
6. Cape May	0.81
7. Camden	0.47
8. Hunterdon	0.45
9. New Castle	0.25

Eight of the 9 counties are in the Delaware Valley. Seven are in or adjacent to southern New Jersey. Given the importance of bladder cancer in the factor, the correlation results parallel those presented under male and female bladder cancer.

The chemical industry, socioeconomic status and the relative presence of German nationals are moderately correlated. When Salem County is deleted, only the socioeconomic status indicator remains significant (Table 91).

TABLE 91

Important Bivariate Correlations: Southern New Jersey Factor

Variable	45 Areas	44 Areas (without Salem County)	40 Areas	39 Areas (without Salem (County)
1. % chemical workers	.317	-	.427*	-
2. % professional workers	-.285	-.262	-.309	-.271
3. % female completed at least college	-	-	-.269	-
4. % German population	-	-	-.255	-
5. TSP wind direction/ distance index	-	-	.253	.315
6. Population density	-	-	-	.272

* Significant at .01 level

In conclusion, with the exception of socioeconomic status, there seems to be little consistency in the pattern of high risk factors that we have available to explain the southern New Jersey factor. The 1968-1972 data which will be presented in Part II offer a tentative explanation.

Northeastern Pennsylvania Female Cancer Mortality Factor

A third factor isolated a group of six counties that exhibited high rates of female kidney, skin, and leukemia cancer and low rates of lung cancer. The unrotated factor loadings and factors scores are summarized in Tables 92 and 93. (The rotated and unrotated solutions are found in Chapter 2).

TABLE 92

Northeastern Pennsylvania Cancer Factor

Variable	Unrotated Factor Loading
1. Female kidney	.694
2. Female skin	.485
3. Female lung	-.480
4. Female leukemia	.442

TABLE 93

Factor Scores for Counties Strongly Identifying with the Northeastern Pennsylvania Cancer Factor

County		Unrotated Factor Score
1.	Sussex	3.13
2.	Sullivan	2.21
3.	Carbon	1.84
4.	Pike	1.74
5.	Hunterdon	1.06
6.	Warren	0.74

The bivariate correlations suggest that the populations of these counties are relatively stable, are low socioeconomic status, and that relatively few are Italians (Table 92).

TABLE 94

Important Bivariate Correlations: Northeastern Pennsylvania Cancer Factor

Variable		45 Areas R
1.	% same house	.476*
2.	% Italian	-.352
3.	% professional workers	-.288
4.	% female completed at least college	-.248

* Significant at .01 level

As in the case of the factor, the explanatory power of the high risk factors is weak and relationships are questionable.

Summary

Recapitulating the results of thousands of mathematical correlations and a large number of hypotheses necessitates generalizations. This section overviews the results from three perspectives: (1) high risk factors; (2) body sites; and (3) counties. Table 95 summarizes the statistical results by body site.

TABLE 95

Summary of Statistical Results by Body Site

Category	Regional Average USA Average	No. significant R values at .05 level	Explanatory Strength Individual High Risk Factors	 Factor Analysis Variables	High and Low Risk Factors
			Mult. R	Mult. R	
1. wm prostate, 177	.99	2 of 171	ns(n=45) to .381(n=21)	-	none
2a. wf bladder, 181	1.18	2 of 174	ns(n=45) to .389 (n=40)	-	German
2b. wm bladder, 181	1.36	13 of 171	.353 (n=44) to .713(n=21)	-	chemicals, socioeconomic status, Italian
3a. wf skin, 190	.95	8 of 174	ns(n=45) to .519(n=21)	-	socioeconomic status agriculture
3b. wm skin, 190	.95	9 of 171	.595(n=45 to .794(n=21)	-	chemicals, German
4a. wf leukemia, 204	.99	5 of 174	ns(n=21) to .345(n=45)	-	Jewish
4b. wm leukemia, 204	.97	42 of 171	.485(n=21) to .540(n=40)	-	Jewish, general urban/industrial
5a. wf kidney, 180	1.04	5 of 174	ns(n=21) to .335(n=40)	-	Polish?, water?
5b. wm kideny, 180	1.10	59 of 171	.494(n=45) to .862(n=21)	-	Polish, USSR, air quality?, socioeconomic status?
6a. wf pancreas, 157	1.07	28 of 174	.492(n=45) to .618(n=21)	.265	socioeconomic status?, air quality?

TABLE 95 (continued)

Summary of Statistical Results by Body Site

Category	Regional Average USA Average	No. significant R values at .05 level	Explanatory Strength Individual High Risk Factors Mult. R	 Factor Analysis Variables Mult. R	High and Low Risk Factors
6b. wm pancreas, 157	1.04	41 of 171	.439(n=21) to .543(n=45)	.324	USSR and Polish, Italian?
7. wf breast, 170	1.13	137 of 174	.550(n=45) to .839(n=21)	. 692	socioeconomic status, foreign born?, local ambient and occupational environments?
8a. wf lung, 162	1.10	30 of 174	.375(n=21) to .462(n=45)	.383	air quality, Italian? female married (smoking?)
8b. wm lung, 162	1.12	126 of 171	.684(n=45) to .785(n=21)	.612	air quality, chemical & transportation equipment workers, Polish, Italian
9a. wf large intestine, 153	1.24	41 of 174	.301(n=40) to .423(n=21)	.317	general urban/industrial, water supply?, family status?
9b. wm large intestine, 153	1.23	133 of 171	.573(n=21) to .639(n=45)	.566	Italian, air pollution? manufacturing?, water supply?
10.a. wf liver, 155	1.03	49 of 174	.524(n=45) to .692(n=21)	.459	socioeconomic status

TABLE 95 (continued)

Summary of Statistical Results by Body Site

Category	Regional Average	No. significant R values at .05 level	Explanatory Strength		High and Low Risk Factors
	USA Average		Individual High Risk Factors Mult. R	Factor Analysis Variables Mult. R	
10.b. wm liver, 155	1.01	129 of 171	.641(n=45) to .747(n=21)	.604	chemical, Russian, Italian air pollution
11a. wf rectum, 154	1.28	122 of 174	.612(n=45) to .670(n=21)	.529	ethnic and urban/industrialization
11b. wm rectum, 154	1.36	130 of 171	.574(n=45) to .694(n=21	.595	ethnic and urban/industrialization
12a. wf stomach, 151	1.13	140 of 174	.768(n=45) to .907(n=21)	.709	ethnic, especially Polish
12b. wm stomach, 151	1.12	140 of 171	.780(n=45) to .859(n=21)	.786	ethnic, especially Polish
13a. wf all, 300	1.10	-	.587(n=45) to .803(n=21)	.530	general urban/industrial
13b. wm all, 300	1.10	-	.785(n=45) to .886(n=21)	.772	general urban/industrial
14. urban/industrial cancer corridor[a]	-	-	.749(n=45)	.737	general urban/industrial
15. southern New Jersey factor[a]	-	-	.599(n=40)	-	chemical, socioeconomic status
16. northeastern Pennsylvania factor[a]	-	-	.476(n=45)	-	socioeconomic status, Italian

ns = not significant

[a] These are factor analysis variables

In general, the personal environmental indicators associated as expected. Socioeconomic status was expected to correlate with cancer of the lung, stomach, liver, bladder, female breast, and perhaps leukemia. Only in the case of lung and leukemia was socioeconomic status not an important indicator. Socioeconomic status appeared in the cases of skin, kidney, and pancreas cancer. We do not know, but we suspect, that in these cases socioeconomic status may be acting as a surrogate for other factors.

The ethnic groups strongly correlated with some of the body sites. Most of these (bladder, leukemia, kidney, lung, large intestine, liver and stomach) had been expected. A few (skin and female breast) were unexpected, particularly female breast. In comparison to each other, the Italian groups demonstrated higher correlations that expected, the USSR lower. In subsequent studies we shall try to develop and incorporate smoking, alcohol, some drug data, and other ethnic groups into the analysis.

The occupational indicators produced mixed results. The general manufacturing category associated with many body site cancers. The metal and construction industries were not as strong as expected. Using all manufacturing as a control variable, we were rarely able to separate their effects from all manufacturing. Metals and construction may be important factors; this study cannot, however, support or reject that contention. On the other hand, the chemical industry produced some of the expected associations. We had expected strong correlations with white male bladder, skin, lung, liver, pancreas, and leukemia cancer. The associations with bladder and skin cancer seem the strongest. The bladder cancer association disappears when Salem County is removed from the data set and then reappears when other counties are deleted. The correlations with liver and lung cancer are strong. However, since the high rate counties are in the urban/industrial core and in counties with high risk ethnic groups, it is possible that chemicals are acting as a surrogate for other high risk factors found in the same counties. Further studies will include additional occupational categories: leather tanning, electrical machinery, scientific and related instruments and perhaps others. Finer scale SIC code and industrial survey data will be added as they become available.

The local ambient environment variables appeared for many body sites. However, it is difficult to separate their effects from occupational and ethnic factors.

Using the number of mortalities and the region's mortality rate compared to the nation's as criteria, the most important body sites are stomach, large intestine, rectum, lung, breast, and bladder. The stomach cancer results are strong. Variation in the rates seem to be strongly tied to ethnic and, in turn, probably dietary factors. The large intestine and rectum results are satisfactory when viewed from the perspective of the explanatory power of the high risk factors. However, the high risk factors are not independent. Thus, the precise roles of different high risk factors cannot be specified. The lung cancer results are not as strong and definitive as the stomach cancer results, but are much clearer than the large intestine and rectum results.

The female breast and bladder analyses are among the most interesting. The female breast results include unexpected local ambient environment and ethnic factors and therefore deserve considerable follow-up. The region's high bladder cancer rates are commonly said to be due to industrial exposure. Therefore, we were not surprised when we found a strong positive association with the relative presence of the chemical industry. Upon further analysis, this association was found to be inconsistent. Therefore, given our present data, the region's high bladder rates cannot be a clearly attributed to a simple set of factors.

Each of the 49 counties in the study is interesting to study because of its rates in one or more diseases. In general, the most important high rate counties are Hudson, Middlesex, and Philadelphia. These three had the highest white male rates and strongly identified with the high risk factors. At the other end of the spectrum Dutchess, Chester, and Morris Counties had relatively low rates.

Some final cautions must be sounded about the results of this study. In addition to the many caveats about data and methodological limitations noted throughout the volume, the reader must recognize that this report reflects only slightly more than one year of study. During the second year, additional high and low risk factors will be added, more recent cancer mortality statistics will be studied, and our increased sensitivity to the melding of the methods and the data will certainly lead us to more useful findings. In short, the relative strengths, of associations reported in this first study, while suggestive of diseases to study and places to study them must not be interpreted as conclusive.

CITATIONS

Barclay, G., Techniques of Population Analysis, New York: Wiley, 1958.

Cleveland W., B. Kleiner, J. McRae and J. Warner, "The Analysis of Ground-Level Ozone Data from New Jersey, New York, Connecticut and Massachusetts: Transport from the New York City Metropolitan Area," Murray Hill, New Jersey: Bell Labs, 1976.

Connecticut Department of Environmental Protection, Connecticut Air Quality Summary-1975, Hartford: Connecticut Department of Environmental Protection.

DeMaio, L. and M. Corn, "Polynuclear Aromatic Hydrocarbons Associated with Particulates in Pittsburgh Air," Journal Air Pollution Control Association, Vol. 16, 1966, pp. 67-71.

Department of Air Resources, City of New York, Data Report-Aerometric Network - Calendar Year 1973, New York: City of New York, Department of Air Resources, Bureau of Technical Services, 1974.

_______., Data Report - Aerometric Network - Calendar Year 1974, New York: City of New York, Department of Air Resources, Bureau of Technical Services, 1975.

_______., Data Report - Aerometric Network - Calendar Year 1975, New York: City of New York, Department of Air Resources, Environmental Protection Administration, 1976.

Department of Environmental Resources, Commonwealth of Pennsylvania, 1975 Pennsylvania Air Quality Surveillance System, Air Quality Report, Department of Environmental Resources, Commonwealth of Pennsylvania, 1976.

Department of Natural Resources, State of Delaware, Monthly Air Quality Data Summary, July 1976, Dover: State of Delaware, Department of Natural Resources and Environmental Control, Division of Environmental Control, 1976.

Department of Public Health, City of Philadelphia, Emissions Inventory and Air Quality Data Report to the Air Pollution Control Board, Philadelphia: City of Philadelphia, Department of Public Health, Air management Services, 1977.

Energy Policy Staff, Office of Science and Technology, Considerations Affecting Steam Power Plant Site Selection, Washington, D.C.: U.S. Government Printing Office, 1968.

Federal Power Commission, Steam-Electric Plant Construction Cost and Annual Production Expenses, 1938 to 1947, Washington, D.C.: U.S. Government Printing Office, Annual reports beginning in 1938.

Freeman, H., S. Levine, and L. Reeder, Handbook of Medical Sociology, Englewood Cliffs, New Jersey: Prentice-Hall, 1972.

Harrison, L., Conversations and written comments, 1978.

LaBelle, C, "Synergistic Effects of Aerosols: Particulates as Carriers of Toxic Vapors", Archives Industrial Health, Vol. 11, 1955, pp. 297-304.

Landsberg, J., "Man-Made Climatic Changes", Science, Vol. 170, 1970, pp. 1265-1273.

Lee, R., "The Size of Suspended Particulate Matter in Air", Science, Vol. 178, 1972, pp. 567-575.

Mason, T. and F. McKay, U.S. Cancer Mortality by County, 1950-69, DHEW Publication No. (NIH) 74-615, Washington, D.C.: U.S. Government Printing Office, 1974.

National Institute for Occupational Safety and Health, Development of an Engineering Control Research and Development Plan for Carcinogenic Materials, Cincinnati: The Agency, 1977.

Natusch, D. and J. Wallace, "Urban Aerosol Toxicity: The Influence of Particle Size," Science, Vol. 186, 1974, pp. 695-699.

New Jersey Department of Environmental Protection, Air Quality Trend Report, Trenton, New Jersey: Department of Environmental Protection, 1976.

New York State Department of Environmental Conservation, New York State Air Quality Report Continuous and Manual Air Monitoring Systems - Annual 1975, Albany: N.Y. State Department of Environmental Conservation, Division of Air Resources, 1976.

Report of the Air Resources Task Group, The Hudson Basin Project, Poughkeepsie, N.Y.: The Hudson Basin Project, 1974.

United States Bureau of the Census, U.S. Census of Population, various volumes, Washington, D.C.: U.S. Government Printing Office, 1950, 1960, 1970.

_______., County Business Patterns, various volumes, Washington, D.C.: U.S. Government Printing Office, 1948, 1959, 1967, 1971.

_______., Census of Manufactures, various volumes, Washington, D.C.: U.S. Government Printing Office, 1947, 1958, 1967, 1972.

United States Environmental Protection Agency, Air Quality Data for 1967 from the National Air Surveillance Networks Revised 1971, Office of Air Programs Publication No. APTD 69-22, Research Triangle Park, North Carolina, 1971.

_______., Air Quality Data for 1968 from the National Air Surveillance Networks, Office of Air Programs, Research Triangle Park, North Carolina, 1972.

_______., Analysis of Population Exposure to Air Pollution in New York-New Jersey-Connecticut Tri-State Region, Environmental Protection Agency, Pub. No. EPA - 450/3 - 76 - 027, Research Triangle Park, North Carolina, 1976.

United States Public Health Service, Air Pollution Measurement of the National Air Sampling Network, Analyses of Suspended Particulates 1953-1957, Publication No. 637, U.S. Government Printing Office, Washington, D.C., 1958.

_______., Air Pollution Measurements of the National Air Sampling Network Analyses of Suspended Particulate Trends 1957-1961, Public Health Service Publication No. 978, U. S. Government Printing Office, Washington, D.C., 1962.

Walbott, G., Health Effects of Environmental Pollutants, St. Louis: C. V. Mosby, 1973.

Water Supply and Pollution Control Division, U.S. Department of Health, Education and Welfare, 1963 Inventory - Municipal Facilities, Washington, D.C.: U.S. Government Printing Office, 1964.

APPENDIX A

STATISTICAL INFORMATION ABOUT THE WHITE POPULATION, 45 UNIT STUDY AREA

1. Means and Standard Deviations

		BODY SITE, ICD CODE	MEAN	STANDARD DEVIATION
1	Male Stomach	151	17.09	2.52
2	Female Stomach	151	8.72	1.49
3	Male Large Intestine	153	20.33	2.85
4	Female Large Intestine	153	20.09	1.89
5	Male Rectum	154	10.41	1.65
6	Female Rectum	154	6.15	0.92
7	Male Liver	155	5.18	0.86
8	Female Liver	155	5.52	0.93
9	Male Pancreas	157	10.01	1.23
10	Female Pancreas	157	6.21	0.79
11	Male Lung	162,163	42.67	5.66
12	Female Lung	162,163	6.92	1.10
13	Female Breast	170	28.82	3.03
14	Male Prostate	177	17.71	1.76
15	Male Kidney	180	4.23	0.57
16	Female Kidney	180	2.07	0.54
17	Male Bladder	181	9.24	1.66
18	Female Bladder	181	2.83	0.57
19	Male Skin	190,191	2.91	0.51
20	Female Skin	190,191	1.72	0.43
21	Male Leukemia	204	8.55	0.79
22	Female Leukemia	204	5.69	0.78
23	Male All	All	192.11	15.35
24	Female All	All	142.78	8.89

2. **Bivariate Correlations Significant at .05 Level**

	VARIABLE	1	2	3	4	5	6	7	8
1	Male Stomach	*	.835	.495	-	.522	.513	.629	.415
2	Female Stomach	.835	*	.505	-	.645	.498	.551	.400
3	Male Large Intestine	.495	.505	*	-	.468	.432	.592	.250
4	Female Large Intestine	-	-	-	*	-	-	.289	-
5	Male Rectum	.522	.645	.468	-	*	.638	.443	-
6	Female Rectum	.513	.498	.432	-	.638	*	.569	.417
7	Male Liver	.629	.551	.592	.289	.443	.569	*	.280
8	Female Liver	.415	.400	.250	-	-	.417	.280	*
9	Male Pancreas	.430	.264	-	.254	.294	-	-	-
10	Female Pancreas	-	.314	.259	-	.384	-	-	-
11	Male Lung	.472	.433	.597	.417	.312	.364	.500	-
12	Female Lung	-	.282	.460	-	-	.390	.348	-
13	Female Breast	.295	.418	.510	.257	.401	.448	.311	-
14	Male Prostate	-	-	-	-	-	-	-	-
15	Male Kidney	.374	.524	-	.298	.311	-	-	.268
16	Female Kidney	-	.252	-	-	.352	-	-	-
17	Male Bladder	-	-	-	.350	-	-	-	-
18	Female Bladder	-	-	-	.479	-	-	-	-
19	Male Skin	-	-	-	-	-	-	-	-
20	Female Skin	-	-	-	-	-	-	-	-
21	Male Leukemia	-	.256	-	.474	-	-	-	-
22	Female Leukemia	-	.277	.262	-	-	.339	-	.446
23	Male All	.750	.708	.789	.427	.640	.552	.691	.249
24	Female All	.545	.639	.548	.529	.621	.567	.504	.380
No. Significant at .05 Level		12	17	13	10	13	12	12	9
No. $\geq$ 0.5		6	7	6	1	5	5	7	0

*To read variable's correlations read across a row or down a column.

2. Bivariate Correlations Significant at .05 Level (continued)

	VARIABLE	9	10	11	12	13	14	15	16
1	Male Stomach	.430	-	.472	-	.295	-	.374	-
2	Female Stomach	.264	.314	.433	.282	.418	-	.524	.252
3	Male Large Intestine	-	.259	.597	.460	.510	-	-	-
4	Female Large Intestine	.254	-	.417	-	.257	-	.298	-
5	Male Rectum	.294	.384	.312	-	.401	-	.311	.352
6	Female Rectum	-	-	.364	.390	.448	-	-	-
7	Male Liver	-	-	.500	.348	.311	-	-	-
8	Female Liver	-	-	-	-	-	-	.268	-
9	Male Pancreas	*	.314	-	-	.255	-	.308	-
10	Female Pancreas	.314	*	-	-	.305	-	.249	-
11	Male Lung	-	-	*	.571	.488	-	-	-
12	Female Lung	-	-	.571	*	.520	-	-	-
13	Female Breast	.255	.305	.488	.520	*	.277	.392	-
14	Male Prostate	-	-	-	-	.277	*	.256	-
15	Male Kidney	.308	.249	-	-	.392	.256	*	-
16	Female Kidney	-	-	-	-	-	-	-	*
17	Male Bladder	-	-	-	-	-	-	-	-
18	Female Bladder	.399	-	.248	-	-	-	-	-
19	Male Skin	-	-	-	-	-	-	-	-
20	Female Skin	-	-	-	-	-	-	-	.438
21	Male Leukemia	.445	-	-	-	.468	-	.441	-
22	Female Leukemia	-	-	-	-	-	-	-	-
23	Male All	.435	.335	.802	.445	.592	.269	.343	-
24	Female All	-	.456	.523	.511	.634	-	.558	.321
No.	Significant at .05 Level	10	8	12	8	16	3	12	4
No.	≥ 0.5	0	0	5	3	4	0	2	0

*To read variable's correlations read across a row or down a column.

2. Bivariate Correlations Significant at .05 Level (continued)

	VARIABLE	17	18	19	20	21	22	23	24
1	Male Stomach	-	-	-	-	-	-	.750	.545
2	Female Stomach	-	-	-	-	.256	.277	.708	.639
3	Male Large Intestine	-	-	-	-	-	.262	.789	.548
4	Female Large Intestine	.350	.479	-	-	.474	-	.427	.529
5	Male Rectum	-	-	-	-	-	-	.640	.621
6	Female Rectum	-	-	-	-	-	.339	.552	.567
7	Male Liver	-	-	-	-	-	-	.691	.504
8	Female Liver	-	-	-	-	-	.446	.249	.380
9	Male Pancreas	-	.399	-	-	.445	-	.435	-
10	Female Pancreas	-	-	-	-	-	-	.335	.456
11	Male Lung	-	.248	-	-	-	-	.802	.523
12	Female Lung	-	-	-	-	-	-	.445	.511
13	Female Breast	-	-	-	-	.468	-	.592	.634
14	Male Prostate	-	-	-	-	-	-	.269	-
15	Male Kidney	-	-	-	-	.441	-	.343	.558
16	Female Kidney	-	-	-	.438	-	-	-	.321
17	Male Bladder	*	.483	.264	-	.394	-	-	-
18	Female Bladder	.483	*	-	-	.252	-	-	.290
19	Male Skin	.264	-	*	-	-	-	-	-
20	Female Skin	-	-	-	*	-	.318	-	-
21	Male Leukemia	.394	.252	-	-	*	-	.422	.511
22	Female Leukemia	-	-	-	.318	-	*	-	-
23	Male All	-	-	-	-	.422	-	*	.753
24	Female All	-	.290	-	-	.511	-	.753	*
No. Significant at .05 Level		4	6	1	2	9	5	17	17
No. ≥ 0.5		0	0	0	0	1	0	9	13

*To read variable's correlations read across a row or down a column.

3 and 4. Multiple Correlation and Final Communality

	VARIABLE	MULTIPLE CORRELATION*	FINAL COMMUNALITY[a]
1	Male Stomach	0.89	0.77
2	Female Stomach	0.94	0.88
3	Male Large Intestine	0.88	0.68
4	Female Large Intestine	0.74	0.56
5	Male Rectum	0.85	0.62
6	Female Rectum	0.80	0.65
7	Male Liver	0.78	0.69
8	Female Liver	0.74	0.65
9	Male Pancreas	0.80	0.62
10	Female Pancreas	0.59	0.37
11	Male Lung	0.95	0.82
12	Female Lung	0.80	0.66
13	Female Breast	0.78	0.67
14	Male Prostate	0.71	0.44
15	Male Kidney	0.70	0.63
16	Female Kidney	0.76	0.69
17	Male Bladder	0.65	0.57
18	Female Bladder	0.84	0.71
19	Male Skin	0.53	0.20
20	Female Skin	0.77	0.64
21	Male Leukemia	0.76	0.58
22	Female Leukemia	0.62	0.48
23	Male All	0.98	0.94
24	Female All	0.94	0.89

*Between one body site and all other body sites.

[a] The communality is the proportion of the variation of each disease found to be in common with the factor analysis variables.

APPENDIX B

STATISTICAL INFORMATION ABOUT THE NONWHITE POPULATION, 35 UNIT STUDY AREA

	VARIABLE	MEAN	STANDARD DEVIATION
1	Male Stomach	25.95	6.55
2	Female Stomach	10.23	3.74
3	Male Large Intestine	14.91	5.19
4	Female Large Intestine	16.10	5.76
5	Male Rectum	7.73	3.46
6	Female Rectum	5.67	2.48
7	Male Liver	6.57	2.64
8	Female Liver	5.94	3.47
9	Male Pancreas	11.69	4.64
10	Female Pancreas	6.52	3.80
11	Male Lung	51.76	12.54
12	Female Lung	8.03	3.38
13	Female Breast	24.05	7.95
14	Male Prostate	29.59	8.67
15	Male Kidney	3.60	2.42
16	Female Kidney	1.51	1.20
17	Male Bladder	6.63	3.33
18	Female Bladder	3.45	1.95
19	Male Skin	1.53	1.21
20	Female Skin	0.86	0.93
21	Male Leukemia	6.85	2.40
22	Female Leukemia	4.25	2.00
23	Male All	223.91	38.97
24	Female All	153.03	33.63

2. Bivariate Correlations Significant at .05 Level

	VARIABLE	1	2	3	4	5	6	7	8
1	Male Stomach	*	-	-	-	-	-	-	-
2	Female Stomach	-	*	-	.444	-	-	-	-
3	Male Large Intestine	-	-	*	-	-	.381	.249	-
4	Female Large Intestine	-	.444	-	*	-	-	-	.565
5	Male Rectum	-	-	-	-	*	-	.501	-
6	Female Rectum	-	-	.381	-	-	*	-	-
7	Male Liver	-	.249	-	-	.501	-	*	-
8	Female Liver	-	-	-	.565	-	-	-	*
9	Male Pancreas	.328	-	.306	.333	-	-	-	.377
10	Female Pancreas	-	-	-	.632	-	-	-	.660
11	Male Lung	-	-	-	-	-	-	-	.418
12	Female Lung	-	-	-	-	-	-	-	-
13	Female Breast	-	.432	.329	.318	-	-	-	.449
14	Male Prostate	-	.401	-	.408	.283	-	-	.279
15	Male Kidney	-	-	-	-	.413	-	.360	-
16	Female Kidney	-	-	.342	-	.300	.357	-	-
17	Male Bladder	.434	-	-	-	-	-	-	.546
18	Female Bladder	-	-	.385	-	-	.261	-	-
19	Male Skin	-	.284	-	-	-	-	.325	-
20	Female Skin	-	-	-	-	-	.395	-	-
21	Male Leukemia	-	-	-	-	-	-	-	.445
22	Female Leukemia	-	.400	-	-	-	-	-	-
23	Male All	.316	-	.480	.360	.342	.254	.293	.367
24	Female All	-	.309	-	.580	-	-	-	.796
No. Significant at .05 Level		3	6	7	8	5	5	5	10
No. ≥ 0.5		0	0	0	3	1	0	1	4

*To read variable's correlations read across a row or down a column.

2. Bivariate Correlations Significant at .05 Level (continued)

	VARIABLE	9	10	11	12	13	14	15	16
1	Male Stomach	.328	-	-	-	-	-	-	-
2	Female Stomach	-	-	-	-	.432	.401	-	-
3	Male Large Intestine	.306	-	-	-	.329	-	-	.342
4	Female Large Intestine	.333	.632	-	-	.318	.408	-	-
5	Male Rectum	-	-	-	-	-	.283	.413	.300
6	Female Rectum	-	-	-	-	-	-	-	.357
7	Male Liver	-	-	-	-	-	-	.360	-
8	Female Liver	.377	.660	.418	-	.449	.279	-	-
9	Male Pancreas	*	.417	-	-	.460	-	-	.267
10	Female Pancreas	.417	*	.302	-	.340	-	-	.380
11	Male Lung	-	.302	*	.519	.264	.505	.273	.282
12	Female Lung	-	-	.519	*	-	.444	.383	-
13	Female Breast	.460	.340	.264	-	*	.296	-	.287
14	Male Prostate	-	-	.505	.444	.296	*	-	.357
15	Male Kidney	-	-	.273	.383	-	-	*	-
16	Female Kidney	.267	.380	.282	-	.287	.357	-	*
17	Male Bladder	.405	-	.376	-	.273	-	-	-
18	Female Bladder	-	.298	-	-	.293	.310	-	.517
19	Male Skin	-	-	-	.263	-	-	.436	-
20	Female Skin	.259	-	-	-	-	-	-	-
21	Male Leukemia	.308	.467	-	-	-	-	-	-
22	Female Leukemia	-	-	.356	-	.369	-	-	-
23	Male All	.463	.419	.779	.488	.350	.716	-	.432
24	Female All	.440	.589	.656	.386	.704	.520	-	-
No. Significant at .05 Level		12	10	11	6	14	11	5	10
No. ≥ 0.5		0	3	4	1	1	3	0	1

*To read variable′s correlations read across a row or down a column.

2. Bivariate Correlations Significant at .05 Level (continued)

	VARIABLE	17	18	19	20	21	22	23	24
1	Male Stomach	.434	-	-	-	-	-	.316	-
2	Female Stomach	-	-	.284	-	-	.400	-	.309
3	Male Large Intestine	-	.385	-	-	-	-	.480	-
4	Female Large Intestine	-	-	-	-	-	-	.360	.580
5	Male Rectum	-	-	-	-	-	-	.342	-
6	Female Rectum	-	.261	-	.395	-	-	.254	-
7	Male Liver	-	-	.325	-	-	-	.293	-
8	Female Liver	.546	-	-	-	.445	-	.367	.796
9	Male Pancreas	.405	-	-	.259	.308	-	.463	.440
10	Female Pancreas	-	.298	-	-	.467	-	.419	.589
11	Male Lung	.376	-	-	-	-	.356	.779	.656
12	Female Lung	-	-	.263	-	-	-	.488	.386
13	Female Breast	.273	.293	-	-	-	.369	.350	.704
14	Male Prostate	-	.310	-	-	-	-	.716	.520
15	Male Kidney	-	-	.436	-	-	-	-	-
16	Female Kidney	-	.517	-	-	-	-	.432	-
17	Male Bladder	*	-	-	-	-	-	.407	.467
18	Female Bladder	-	*	-	.358	-	-	.322	-
19	Male Skin	-	-	*	-	-	-	-	-
20	Female Skin	-	.358	-	*	-	-	.333	-
21	Male Leukemia	-	-	-	-	*	-	-	.264
22	Female Leukemia	-	-	-	-	-	*	-	.303
23	Male All	.407	.322	-	.333	-	-	*	.673
24	Female All	.467	-	-	-	.264	.303	.673	*
No. Significant at .05 Level		7	8	4	4	4	4	18	13
No. ≥ 0.5		1	1	0	0	0	0	3	7

*To read variable's correlations read across a row or down a column.

3 and 4. Multiple Correlation and Final Communality

	VARIABLE	MULTIPLE CORRELATION*	FINAL COMMUNALITY[a]
1	Male Stomach	.75	.57
2	Female Stomach	.91	.89
3	Male Large Intestine	.91	.51
4	Female Large Intestine	.88	.60
5	Male Rectum	.75	.56
6	Female Rectum	.87	.62
7	Male Liver	.88	.78
8	Female Liver	.96	.95
9	Male Pancreas	.84	.72
10	Female Pancreas	.98	.85
11	Male Lung	.96	.80
12	Female Lung	.92	.83
13	Female Breast	.95	.70
14	Male Prostate	.93	.67
15	Male Kidney	.91	.81
16	Female Kidney	.90	.62
17	Male Bladder	.93	.79
18	Female Bladder	.77	.61
19	Male Skin	.78	.46
20	Female Skin	.83	.62
21	Male Leukemia	.84	.69
22	Female Leukemia	.92	.85
23	Male All	.99	.97
24	Female All	.99	.91

---

*Between one body site and all other body sites.

[a] The communality is the proportion of the variation of each disease found to be in common with the factor analysis variables.